Fake News Detection Using Traditional and Deep Learning Methods: A Comparative Study on LIAR1 and LIAR2 Datasets

1st Caizhuangzhou Cui
School of CSE
University of New South Wales
Sydney, Australia
z5509061@ad.unsw.edu.au

2nd Hefei Fu
School of CSE
University of New South Wales
Sydney, Australia
z5597779@ad.unsw.edu.au

3rd Tingyuan Liu
School of CSE
University of New South Wales
Sydney, Australia
z5565888@ad.unsw.edu.au

4th Yankun Wei
School of CSE
University of New South Wales
Sydney, Australia
z5528317@ad.unsw.edu.au

5th Zihan Chen
School of CSE
University of New South Wales
Sydney, Australia
z5527498@ad.unsw.edu.au

***Abstract*—A study examines multiple machine learning methods for fake news detection which contains traditional SVMs and recurrent neural networks (LSTM) together with transformer-based models (BERT). The models' performance is measured through evaluation on LIAR1 and LIAR2 datasets by assessing binary (real vs fake) and six-class categorization results. Our research shows that BERT provides superior classification results particularly when processing LIAR2 dataset because it contains elaborate structural and contextual elements. GPT-4 helps us review incorrectly classified examples which shows that annotations need clarification and suggests using LLMs for evaluation framework integration.**

***Keywords*—Fake News Detection, Natural Language Processing, Multi-class Classification, LIAR Dataset, BERT, LSTM, SVM, GPT-4, Large Language Models**

I. Introduction

The fast-expanding dissemination of unconfirmed information and fake news constitutes a crucial obstacle for digital communication systems which affects political discussions along with healthcare messaging and voting initiatives. False information quickly spreads across large audiences because of social media platforms and algorithm-driven content distribution which happens before fact-checkers can verify them.

A computational analysis of fake news detection requires more than syntactic and semantic processing since it involves contextual understanding and establishing legitimate information sources while integrating factual knowledge. Automated systems using natural language processing (NLP) and machine learning techniques now prove essential for finding deceptive or misleading content because of extensive online information quantity and its vague nature.

Research on fact-checking uses the LIAR dataset from Wang et al. [1] as the standard benchmark throughout the field. Researcher volunteers curated over 12000 short political statements that received six-step truthfulness categorizations including false statements among others. The LIAR2 expansion added structural metadata features to speaker information and party affiliation together with credibility history and subject tags which allowed researchers to build better modeling solutions.

This research solution handles the challenge of six-class multi-classification together with binary fake vs. real classification for both LIAR1 and LIAR2 datasets. The performance of three representative model families receives assessment in this work.

The first traditional approach connects TF-IDF feature extraction methods to a Support Vector Machine (SVM) classifiers.

An LSTM neural network handles the sequence-based dependencies found in textual data.

The BERT-based transformer model acts as a state-of-the-art performer across diverse NLP benchmark tests.

We perform thorough exploratory data analysis (EDA) to identify distributional biases along with label imbalances while investigating patterns that depend on speakers and subjects which affect model responses. The implementation of GPT-4 as an extra annotation system evaluates selected BERT misrecognitions through re-assessment. The procedure enables examination of genuine annotation reliability while testing whether LLMs could enhance typical classification algorithms.

Research explores fake news detection performances of different models including transformer-based ones and presents methods to enhance robustness and interpretability through LLMs.

II. Literature Review

A. Traditional Methods

The earliest methods for fake news detection relied on manual features which included n-grams alongside term frequency-inverse document frequency (TF-IDF) analytics and emotion indicators. SVMs and Logistic Regression operated as strong baseline models because their low computational demands combined with the ease of interpretation [1] [2]. The approach developed by Potthast et al. [3] confirmed the usefulness of stylometric features

together with content complexity indicators to detect fake and real news articles in their research. The classification models demonstrate effectiveness on regular assignments yet face difficulties in complex generalization tasks featuring uncertain semantic meanings especially when the class boundaries remain unclear (for instance barely-true and half-true judgments).

B. .Deep Learning

Deep neural networks made it possible for systems to learn end-to-end directly from raw text without depending on manual feature engineering steps. Long Short-Term Memory (LSTM) models lowered their popularity among Recurrent Neural Networks (RNNs) which wanted to model sequential word dependencies [4]. The processing of fake news detection accuracy has proven higher when LSTM models use attention techniques and external metadata according to research [5]. The short input text in LIAR hinders performance benefits of sequential context models as used in detection systems.

C. Transformers and BERT

NLP experienced a significant change when Transformers and pre-trained models like Bidirectional Encoder Representations from Transformers (BERT) were introduced in [6]. The advanced bidirectional attention mechanism coupled with masked language modeling that BERT implements provides it with exceptional performance for classifying sentences including fake news identification. Studies show BERT-based models achieve superior performance than traditional classifiers alongside RNNs when processing the LIAR dataset [7] [8] especially when identifying fake versus real content during binary operations because the distinction becomes more distinct. BERT offers a fine-tuning capability that enables it to adjust according to the target task distribution.

D. Recent Advances : LLMs and Weak Supervision

The fact-checking research field expanded because of modern developments in large-scale language models including GPT-3 and GPT-4. The models function in three settings from zero-shot to few-shot to prompt-tuned evaluation for claim truth detection without specialist training [9]. Research investigates the ways LLMs verify classifier outcomes in post-hoc evaluations and apply soft-labels for semi-supervised learning and employ ensemble methods having LLMs function as review partners [10]. Different methods employing data augmentation and label smoothing and focal loss together with adversarial training have emerged to handle both label imbalance issues and model overconfidence [11] [12].

III. METHODS & DATASET ANALYSIS

The overall research methodology breaks down how this study was conducted to include the processing of datasets and the development of model structures and training parameters and evaluation methods for binary and six-class fake news classification.

A. Dataset Preprocessing and Feature Engineering

The research relies on LIAR1 [1] and an extended version LIAR2 that provides additional sample data along with expanded metadata. Both datasets use the statement field as their main input that contains political claims which span between 5 and 50 tokens in length.

The preprocessing methods match different models through their distinct task requirements.

The statement input for SVM adopts a filtered version clean_statement_svm that eliminates stopwords and weak lexemes.

The processing sequence for LSTM begins with standard tokenization on clean_statement for sequences that require padding to maximum length (max_len = 100).

The BERT processing stage accepts clean_statement_bert across its original surface form through HuggingFace BERT tokenization that produces pairs of input_ids and attention_mask.

The labels receive integer values from 0 to 5 which serve as a numeric code that represents:

0 - pants-fire, 1 - false, 2 - barely-true, 3 - half-true, 4 - mostly-true, 5 - true.

We assign the labels [0, 1, 2] to the fake class while labeling [3, 4, 5] with real.

B. Model Architectures

1.Traditional Machine Learning: SVM

The implementation utilizes a Support Vector Machine (SVM) classifier with RBF kernel which receives TF-IDF vectors based on a vocabulary containing 5000 terms. The TF-IDF matrix derives from clean_statement_svm by analyzing both unigram and bigram features.

As a strong non-neural baseline this model delivers both interpretability along with fast training but it lacks semantic generalization ability.

2. Sequence Model: LSTM

Our LSTM model consists of:

An embedding layer (initialized randomly),

An LSTM network with a single layer contains 128 hidden units.

A final layer receives as input the hidden state for projecting it into 6 final output clusters.

The optimization strategy includes Adam with learning rate set to 1e-3 while dropout regularization uses an intensity of 0.3. Each input sequence extends to reach a total of 100 tokens.

3. Transformer Model: BERT

The HuggingFace Transformers platform provides us with its bert-base-uncased model. The architecture includes:

BERT encoder with 12 layers and 110M parameters,

A classification head: linear layer + softmax over 6 (or 2) classes.

At the beginning of earlier experiments the final layer received fine-tuning however in later experiments we applied fine-tuning to the entire BERT encoder. Our training utilizes learning rate set to 2e-5 while batch size stands at 16 until early stopping activates through validation loss measurements from 3–5 epochs.

C. Training Setup and Classification Tasks

We address two classification settings:

1. The six-class task requires a direct association between each input statement and one of the available truth labels.
2. The data makes use of a reduction to binary classes which groups statements between fake and real categories.

The datasets were divided into three parts using the following proportions:

80% training, 10% validation, 10% testing,

The method of stratified sampling enables preservation of the class distribution between groups.

The testing of all models extends to both LIAR1 and LIAR2 datasets (when possible) to verify generalization between different datasets. Leaders with high activity levels (including Trump, Clinton, and Obama) represent most entries in the database. The distribution of labels shows Republicans dominating false or pants-fire statements yet Democrats distributing more toward mostly-true and half-true assessments. The discovered relationships indicate possible problems with biased labels as well as the danger of models that become too specialized in recognizing speakers.

The research data shows that some specific topics including healthcare and federal budget produce overwhelmingly high numbers of fake news cases throughout the study. Politically prominent states (e.g., Texas, California) account for more samples, likely reflecting media coverage skew rather than actual state-wise news trends.

D. Evaluation Metrics

The evaluation uses standard multi-class and binary classification metrics together with accuracy scores along with precision rates and F1-scores (with macro-averaging) and confusion matrices to display class-specific mistakes.

Accuracy: Overall correctness.

Precision, Recall, F1-score (macro-averaged).

Confusion Matrix: To visualize class-specific confusion patterns.

Our evaluation includes F1-score metrics individualized for fake and real category classifications. During six-class operations we use macro-averaged F1-score because labels do not distribute evenly.

E. LLM-Based Label Reassessment

GPT-4 serves for post-hoc label validation through an experimental process which helps quantify ground-truth reliability and identify classification errors.

We retrieve from the test set all samples which BERT incorrectly labeled.

The statement and possible label definitions are provided to GPT-4 for processing every test sample.

GPT-4 records its selected prediction along with explanation data for assessment purposes.

Evaluation through this process reveals the cause of model errors along with meaningful insights about uncertain cases.

F. Error Case Sampling

A qualitative assessment of prediction errors served as the method for determining interpretability across models. We reviewed select cases with forecast errors to determine if they resulted from model boundaries or ambiguous labels.

Each model SVM together with LSTM and BERT received our attention as we analyzed their highest misidentified samples from both binary classification and the six-class classification scenario. The analysis showed different detection patterns for misclassifications that included keyword discrimination in SVM and LSTM sensitivity to polarity along with label uncertainties in BERT.

These illustrations demonstrated the manner in which models are influenced by different text characteristics together with sentence organization as well as semantic complexity. Truthfulness labels often showed faint distinctions between half-true and mostly-true among other categories which made human labeling particularly challenging. Model failure evaluation depends not only on performance tests but also on understanding how models breakdown and interpretation of such breakdowns for better data assessment.

IV. Experimental Results

We evaluate three distinct classification models to detect fake news among which we include Traditional machine learning (TF-IDF + SVM) Deep learning using LSTM and Transformer-based model using BERT.

Traditional machine learning (TF-IDF + SVM)

Deep learning using LSTM

Transformer-based model using BERT

The classification tasks encompass six-class and binary objectives that use LIAR1 and LIAR2 datasets for evaluation.

A. Overall Accuracy Comparison

Model	Liar1		Liar2	
	Six	Binary	Six	Binary
SVM	0.24	0.60	0.31	0.66
LSTM	0.24	0.59	0.31	0.68
BERT	0.27	0.63	0.35	0.70

Table 1

B. Classification Report (Selected Metrics)

Below are selected class-level results for Liar2 comparing model precision, recall, and F1-score:

Model	Binary				Six-class			
	Precision	Recall	F1-score	acc	Precision	Recall	F1-score	acc
SVM	0.29	0.25	0.25	0.66	0.65	0.65	0.65	0.31
LSTM	0.30	0.28	0.27	0.68	0.68	0.68	0.68	0.31
BERT	0.33	0.30	0.28	0.70	0.70	0.70	0.70	0.35

Table 2

C. Confusion Matrix

A complete review of confusion matrices exists for all evaluated models to expose common classification errors.

The majority of SVM (Six-class LIAR2 model targets false labels more often than detecting pants-fire or true statements.

LSTM (six-class, LIAR1): High confusion between barely-true, half-true, and false; weak diagonal structure.

BERT (six-class, LIAR2): Improved diagonal concentration, but low recall for true and mostly-true.

The study conducted by GPT-4 (post-hoc label rejudegment) demonstrated that BERT's incorrect classifications may stem from actual inconsistencies in the ground truth labels.

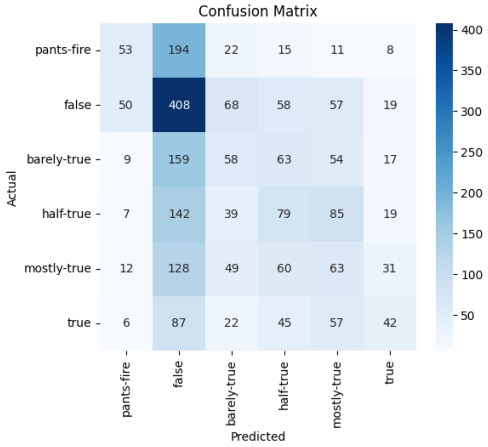

Figure 1. Confusion Matrix – SVM

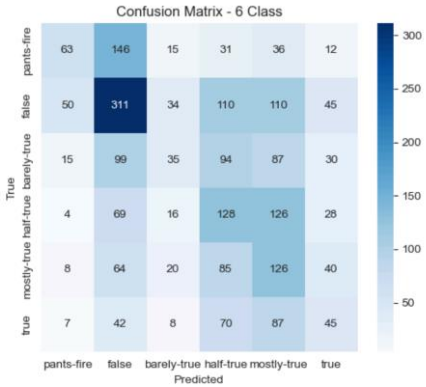

Figure 2. Confusion Matrix – LSTM

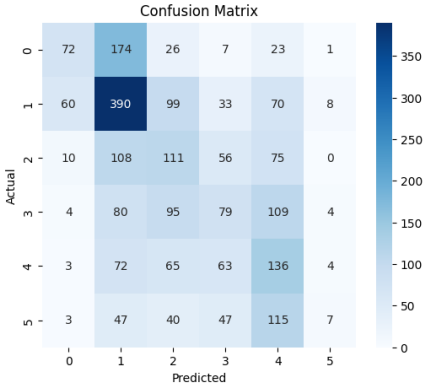

Figure 3. Confusion Matrix – BERT

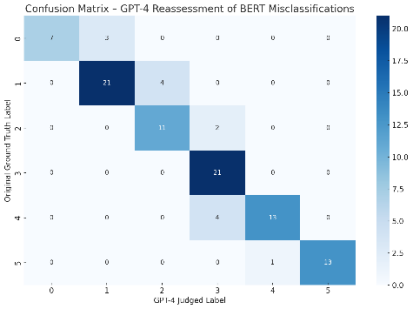

Figure 4. Confusion Matrix – GPT-4o

D. *Training Curves (DL Models Only)*

Within LSTM (LIAR2, six-class) the training loss steeply declines then validation accuracy reaches a sustained level of approximately 0.25.

BERT (LIAR2, binary) achieves smooth convergence and validation accuracy steadily increases to around 0.70.

The validation loss of BERT (LIAR2 using six classes) shows steady improvement while six-class accuracy remains stagnant because of diagnostic problems.

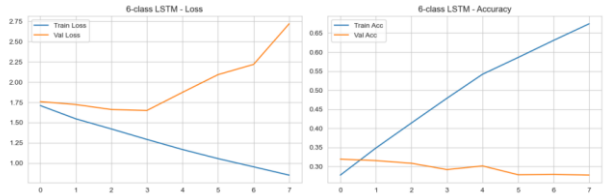

Figure 5. LSTM Accuracy & Loss vs Epoch

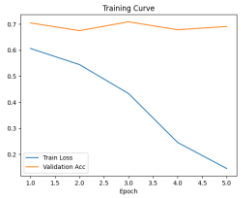

Figure 6. BERT Accuracy & Loss vs Epoch (Binary)

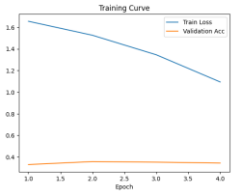

Figure7. BERT Accuracy & Loss vs Epoch (Six-class)

E. *Qualitative Examples: GPT-4 Post-Hoc Analysis)*

We tested 100 test samples that BERT misidentified by asking GPT-4 to review them again. GPT-4's semantic judgment matched:

The original label in 67% of cases

The BERT prediction in 33% of cases

Labeling inconsistencies within the dataset help explain some errors that occur during classification processes. GPT-4 delivered explanations that pointed out sentence wording as well as broad generalizations and context falls.

Judgment Alignment	Count	Percentage
Matches Ground Truth	67	67%
Matches BERT Prediction	33	33%

Table 3

F. *Model Summary Table*

Model	Strengths	Limitations
SVM	Fast training, interpretable	Poor generalization; favors majority class
LSTM	Sequence modeling, modest accuracy	Sensitive to short inputs; weak contextual understanding
BERT	Best performance, strong semantics	Misclassifies fine-grained truth classes; label ambiguity remains

Table 4

G. Representative Misclassified Examples across Models

Model	Task	Statement	True Label	Predicted label
SVM	Binary	"AHCA will significantly reduce insurance premiums."	Real	Fake
SVM	Six	"Joe Biden is a pedophile."	Pants-Fire (0)	False (1)
LSTM	Binary	"America was the only country that ended slavery."	Fake	Real
LSTM	Six	"Hillary Clinton has taken lobbyist money."	True (5)	Mostly-True (4)
BERT	Binary	"The Texas House speaker was thanked by Planned Parenthood."	Real	Fake
BERT	Six	"The Federal Register weighs over 340 pounds."	Half-True (3)	Mostly-True (4)

Table 5

V. DISCUSSION

The research examines three fake news classification methods through SVM, LSTM and BERT while assessment takes place by using LIAR1 and LIAR2 datasets for six-way and binary classification tasks. The analysis demonstrates essential knowledge about task complexity as well as model abilities and restrictions and dataset effects and potential benefits of LLM-based evaluation.

A. Task Complexity: Binary vs. Fine-grained Classification

The results of our experiments support an important observation made previously about the simpler task of binary classification when compared to six-class classification. The model faces a simpler task in binary classification since it requires determining if a statement is factual or not while human perception works in a similar binary manner.

The process of six-class identification demands precise distinction between truth levels that human evaluators might find challenging to judge when comparing barely-true statements to half-true ones and between mostly-true and fully-true declarations. The identification labels show overlapping meanings and require contextual analysis because they lack well-defined boundaries. The grounds for making incorrect classifications stem from ambiguous boundaries defined by the ground-truth rather than from issues with model performance. The existing label system presents structural problems because it demonstrates poor reliability for serving as a supervised learning target.

B. Model Comparison : Performance, Generalization, and Interpretability

The three models exhibit distinct behaviors and capabilities:

SVM operates efficiently and delivers interpretable results through features from TF-IDF which make decisions transparent to users. Performance is confined to surface features of the text because the model lacks the capability to process semantic meaning or pragmatic cues as demonstrated by examples such as sarcasm and negations. The word frequency factor dominates decision-making which results in a strong emphasis on frequency over semantic interpretation.

Across different tasks and datasets LSTM models show a blend of performance results during binary classification on LIAR2 they achieved 68% accuracy which reached BERT's level of performance and exceeded SVM results. The six-class category evaluation produces limited results which amount to 31% accuracy on LIAR2 and 23% accuracy on LIAR1 because of label semantic similarity and uneven class distribution.

The model showed above-average ability to distinguish between false and half-true statements on LIAR1 although it did not utilize deep contextual models. This indicates LSTM can detect systematic patterns when recognizing some recognizable labels.

In LIAR1 and LIAR2 datasets BERT achieves stronger performance than the baselines and demonstrates equivalent results. BERT delivered 63% precision for LIAR1 binary task and 27% accuracy for LIAR1 six-class classification surpassing both SVM and LSTM. The experimental results validate that BERT demonstrates reliable performance across less structured perspectives of the input data. In the six-class setting BERT fails to correctly identify labels while its attention patterns operate as an unexplained mystery preventing practical transparency.

C. Dataset Effects : Scale, Structure, and Bias

Dataset quality stands as a fundamental determinant for model performance because of the difference between LIAR1 and LIAR2.

LIAR1 contains about 12k examples but suffers from class imbalance and redundant speakers across statements and unreliable annotation quality that particularly affects the performance of LSTM and SVM models. The lack of structural metadata prevents the use of auxiliary features because the dataset does not include information such as speaker credibility.

LIAR2 offers users access to more than 23k samples and includes abundant side information that includes speaker party status and historical record documentation plus cleaned statements to optimize model specification. The extra information proves beneficial for transformer-based models to learn more effectively.

The results prove dataset design improvements rely on equal representation of samples and well-defined terminology within structured organizational fields as essential as improving model construction methods.

D. Misclassification Patterns : Structure vs Semantics

The main classification issues identified through error analysis include two primary categories:

Original mistakes in declaration emerged because of proximity indicators between similar labels (such as half-true next to barely-true) and are often due to vague definitions or subjective original annotation approach.

Errors of semantic confusion appear when modeling systems fail to interpret pragmatic indicators about hyperbole and negation and unprovable assertions. The mostly-true and pants-fire categories tend to produce such classification errors.

The label collapsing phenomenon even occurs in BERT models alongside other high-performing models by having them default to false or half-true responses when they are unsure. The middle classes demonstrate ambiguity while the uneven distribution of data points probably plays a significant role in this result.

A deeper analysis of incorrect classifications showed that SVM misidentifies statements that contain words commonly used in political debates or numerical evidence. The presence of Obamacare together with insurance premium and political figures in training data resulted in SVM's incorrect labeling of factual claims. Because it lacks semantic or contextual modeling SVM reduces to high-frequency token to label prior assignments that lead to systematic bias.

The LSTM model detects sequential information but it reacts strongly to certain polarity signals like "never," "only," and "always." The neural network detects the words as signs of extreme statements but acknowledges them as deceptive. This creates an effect where assertive statements in true claims become downgraded by the system. The models demonstrate this tendency most prominently in cases of historical or nationalistic statements because tone dominates over content.

BERT achieves the highest overall accuracy because of its profound bidirectional attention functionality and extensive pretraining on vast text data. BERT faces difficulty distinguishing between barely-true along with half-true and mostly-true categories when performing semantic evaluations. Several times BERT detected the factual organization of a statement correctly but it faltered in the interpretation of its statement strength or generality level because factual scores often overlap in those areas.

The research shows that model deficiencies alone do not explain classification mistakes because truth labels tend to contain ambiguous and blurry definitions. Within six-class settings the labels prove ambiguous because they exist as subjective elements which frequently overlap and depend heavily on context while human annotators also show difficulties in agreement. The situation requires both contextual reasoning and weak supervision as GPT-4 together with other tools can help with post-hoc analysis and label tuning and uncertainty measurement which enhances interpretation quality while ensuring reliability.

E. GPT-4 as a Label Re-Evaluator and Human-AI Collaborator

We purposefully introduced GPT-4 for retrospective evaluation of misidentified samples by BERT to determine if the incorrect classifications held true. The findings were insightful:

GPT-4 confirmed BERT's output as correct when the original label was incorrect for 33% of the examined cases.

The contextual plausibility along with vague language and speaker contextualization potential served GPT-4 as explanations even though these features were inaccessible to human labelers.

The process demonstrates that ground truth values in subjective classification tasks do not automatically emerge from labels. LLMs like GPT-4 can serve as:

The tool acts as a soft labeling instrument to deliver probabilistic information between multiple categories.

Disagreement detectors will identify situations where annotators have conflicting opinions.

The justification engine performs logical explanations about predictions through human-relevant reasoning.

The integration of machine models with LLMs establishes humane collaborative fact-checking operations which identify uncertain classifications for human-LLM dispute resolution methods.

Client analysis revealed certain cases labeled as misclassifications actually showed inconsistencies between annotation standards and semantic meaning interpretation instead of model-related problems. When the model identifies a statement as half-true while human evaluators rate it barely-true it is possible for the statement to fit into both categories based on interpretation. The evaluation through accuracy and F1 scores fails to determine borderline cases accurately therefore it may incorrectly judge model performance on situations with reasonable generalizations.

The evaluation process should include model-ground-truth disagreement evaluation according to our proposed future work. Differing model predictions from ground truth labels should not be considered outright failures since they might warrant further interpretability assessment or human oversight. GPT-4 functions as a key tool for semantic validation alongside classification because it provides explanations with alternative soft-label responses.

F. From Academic Benchmark to Real-World Deployment

The performance accuracy of present-day fake news detectors makes them unavailable for direct use in critical applications. Several hurdles remain:

Users must understand the basis for declaring a statement incorrect for it to build their trust.

The training of models on political claims does not guarantee their ability to recognize health misinformation and financial fraud.

Research into non-English and cross-cultural misinformation has not received enough investigation.

Fake news creators successfully manipulate content to bypass classifiers with special focus on systems based on keywords such as SVM.

A hybrid pipeline that combines:

BERT-style semantic modeling

LLM-assisted validation

Implementing human-in-the-loop annotation will create a solution which combines strength with transparency and scalability.

VI. CONCLUSION

We evaluated an entire framework for fake news detection between LIAR1 and LIAR2 datasets with three model types—SVM, LSTM, and BERT. Our research addressed both real-fake binary decisions and six-class scenarios and enriched quantitative findings with exploratory visualizations

alongside confusion matrix assessments as well as human reading evaluation by GPT-4.

Our research generates multiple essential survey outcomes.

Transformer models based on BERT produce superior results than traditional and sequential baselines when performing binary classification and deliver highest precision-recall balance with optimal accuracy levels. BERT maintained excellent performance in binary LIAR1 classification (63%) while demonstrating the best results (27%) among all models in the six-class category. Under noisy labeling conditions BERT maintains excellent generalization abilities.

Our findings demonstrate that the performance of LSTM reaches an effective level in terms of complexity management with its 68% accuracy achieved in the binary LIAR2 task.

The results indicate that LSTM maintains effectiveness in lower-resource or edge computing applications as long as its parameters are correctly adjusted although transformer models might be unpractical within these scenarios.

The way a dataset is organized through its structure elements affects model generalization because it determines label accuracy combined with balanced classes and detailed features. The updated LIAR2 system brought significant performance advancements compared to the previous version LIAR1.

The difficulty to perform six-class classification persists because bare-true and half-true annotations have ambiguous semantic meanings that make precise categorization challenging. The use of supervised learning methods reveals essential barriers that exist when assessing subjective truth scales.

Because it operates as a large language model GPT-4 functions as an auxiliary evaluator which detects errors in ground-truth annotations while providing human-oriented reclassification explanations.

Research evidence suggests developing fake news detection systems through a combination of statistical learning approaches and contextual language models and LLM-based reasoning methods.

A. Project Contributions

This work brings forward three main contributions.

Modeling Spectrum Exploration

The research performed assessment and comparison between SVM, LSTM and BERT models while testing these algorithms on traditional binary and multi-class problem sets.

Cross-Dataset Performance Benchmarking

The performance evaluation between LIAR1 and LIAR2 datasets demonstrated the significance of data quantity and organization as well as labeling precision to model accuracy.

Misclassification and Bias Diagnosis

A thorough examination of difficult prediction tasks was performed through confusion matrix evaluation and error pattern detection across different labels and topics.

LLM-augmented Label Review

The research team employed GPT-4 to verify BERT's incorrect predictions through label review assessment which validated LLM application value for judgment assistance.

Toward Explainable Pipelines

Comments were made regarding model interpretability requirements for next-generation systems to deliver human-comprehensible explanations when used in public domain applications.

B. Future Work

This work brings forward three main contributions.

Modeling Spectrum Exploration

An analysis of traditional (SVM) and sequential (LSTM) and transformer (BERT) models for both binary and multi-class tasks took place.

Cross-Dataset Performance Benchmarking

The examination of model performance between LIAR1 and LIAR2 datasets revealed the main role played by dataset size and organizational patterns as well as annotation precision.

Misclassification and Bias Diagnosis

A thorough examination of difficult prediction tasks was performed through confusion matrix evaluation and error pattern detection across different labels and topics.

LLM-augmented Label Review

The research team employed GPT-4 to verify BERT's incorrect predictions through label review assessment which validated LLM application value for judgment assistance.

Toward Explainable Pipelines

Comments were made regarding model interpretability requirements for next-generation systems to deliver human-comprehensible explanations when used in public domain applications.

Create a subset of LIAR modified by human annotators together with LLM analysis. Create a manually-reviewed validation segment that employs jointly expert and GPT-4 evaluations to address conflicts between predictions from the model and original truth-based annotations. This subset functions as the benchmarking standard for validating fine semantic comprehension abilities.

Develop explanation-based training objectives The training process should instruct models to determine accurate labels in addition to producing explanations about their selection process along with confirming evidence. Using this method improves the differentiation between fine-grain classes such as barely-true versus half-true since interpretation matters more than the underlying facts.

Use adversarial and contrastive examples The training procedure uses pairs of scarcely different truthful statements to make the model develop robust decision boundaries. Applying this system helps distinguish similar classifications thus leads to more accurate prediction confidence levels.

Incorporate disagreement-aware metrics The evaluation process should adopt new scoring methods such as soft accuracy and confidence-weighted systems and semantic proximity-based penalties for better assessment of models which make nearby incorrect predictions.

This study lays the groundwork for a next-generation fact-checking pipeline that is not only statistically accurate, but also context-aware, explainable, and human-aligned—qualities essential for addressing the growing threat of misinformation in the digital age.

References

[1] Y. Shu, A. Sliva, H. Wang, J. Tang, and B. Liu, “Fake News Detection on Social Media: A Data Mining Perspective,” ACM SIGKDD Explorations, vol. 19, no. 1, pp. 22–36, 2017.

[2] W. Y. Wang, “Liar, Liar Pants on Fire: A New Benchmark Dataset for Fake News Detection,” in *Proc. ACL*, 2017.

[3] M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff, and B. Stein, “A Stylometric Inquiry into Hyperpartisan and Fake News,” in *Proc. ACL*, 2018.

[4] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, “Hierarchical Attention Networks for Document Classification,” in *Proc. NAACL*, 2016.

[5] N. Karimi and H. Tang, “Multi-source Fake News Detection Using Ensemble Learning,” in *Proc. EMNLP*, 2020.

[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.

[7] S. Kaur and K. K. Singh, “Fake News Detection Using Pre-Trained Language Models,” in *Procedia Computer Science*, vol. 199, pp. 754–761, 2022.

[8] A. R. Hanselowski, A. Zhang, P. Zhang, and I. Gurevych, “Fact Check: Assessing the Factual Correctness of Text,” in *Proc. ACL*, 2018.

[9] T. Brown et al., “Language Models are Few-Shot Learners,” in *Proc. NeurIPS*, 2020.

[10] J. Thorne, A. Vlachos, “Evidence-Based Fake News Detection with LLMs,” in *Journal of AI Research*, 2022.

[11] Y. Zhang et al., “Deep Long-Tailed Learning: A Survey,” in *IEEE TPAMI*, vol. 45, no. 9, pp. 10795–10816, 2023.

[12] D. Lin, S. S. Keerthy, and A. Wu, “Adversarial Training for Robust Fake News Detection,” in *Proc. AAAI*, 2021.