Combating Misinformation: A Comparative Evaluation of Zero-Shot Models for Fake News Detection

Sajin Mahmud Arpon, *Member, IEEE*
Department of Computer Science
American International
University-Bangladesh
22-46629-1@student.aiub.edu

Md Ariful Islam
Department of Computer Science
American International
University-Bangladesh
22-46139-1@student.aiub.edu

Mir Md Mofakkar Hossain
Department of Computer Science
American International
University-Bangladesh
22-46245-1@student.aiub.edu

Md Saef Ullah Miah
Department of Computer Science
American International University-Bangladesh
safe@aiub.edu

***Abstract*—The number of fake contents being spread on the internet platforms in rapid succession has led to the importance of the detection of fake news. One practical step is zero-shot classification with pre-trained language models that do not need further fine-tuning. In this paper, four transformer models DeBERTa, BART, ModernBERT, and CE-DeBERTa are evaluated on a smaller variant of the Kaggle Fake News dataset licensed under CC BY-4.0. Majority Voting and Soft Voting (with 0.85 and 0.92 threshold) are also being used as ensemble learning method to enhance reliability. The top-accuracy (76.91%), and F1-score (72.64%), and the greatest degree of uncertainty (59.40) were found in DeBERTa. However, ModernBERT showed less performance (F1-score: 62.40%) with the lowest uncertainty (31.09%). DeBERTa achieved the highest score of 0.6939 based on the composite performance score. The highest accuracy (70.23%) and the highest uncertainty (31.99%) were obtained with Soft Voting with a threshold of 0.92 among ensemble methods. But Majority Voting turned out to have the strongest overall performance score (0.7089). In general, DeBERTa and Majority Voting ensemble are rather successful in terms of trustworthy fake news identification.**

***Index Terms*—Fake News Detection; Misinformation; Zero-Shot Learning; Transformer Models; Natural Language Processing; Ensemble Learning; Text Classification;**

I. Introduction

The spread of fake news is a major issue in the current state of information. Supporters of the ease of creating content and sharing it easily on social media platforms have suggested the spread of misinformation can easily reach large numbers of people, play into the hands of those supporting hate, division, weakening of democratic institutions, and public opinion. Since online discussion is highly dynamic, the mechanisms required to detect and counter fake news should be highly scalable, adaptable, and accurate in order to meet the demand.

Conventional methods of fake news identification largely make use of supervised machine learning models, i.e., the learning of classifiers based on large-scale, label-intensive data in order to sort between legit and illegitimate content [1], [2]. Such techniques, although effective in a managed environment, have serious structural deficiencies associated with them. Obtaining well annotated datasets is not only resource and time-consuming, but also unrealistic in domains of emerging or fast changing misinformation. Also, trained models will have significant generalization gaps when moving to areas or issues that are not covered in the training corpus and will therefore be vulnerable to any new versions of false news.

In order to overcome these difficulties and the concerns of preserving scalability Zero-Shot Learning (ZSL) has become one of the more popular methods in the sphere of Natural Language Processing (NLP). ZSL by design allows classifying unlabelled categories of labels or data domains without using task-specific training information. This has become possible by reliance on massive pre-trained language models; these models encapsulate transfers of previous linguistic and semantic knowledge. Recent work has shown that ZSL can be used to powerful effect in text classification by encoding the inference problem as Natural Language Inference (NLI), or by using prompt-based learning. These models have a considerable ability to quickly adapt new misinformation tropes, without making searches on them to label data expensively, which makes them specially well suited to being matched to real-time fake news detection.

This study introduces a robust evaluation platform for zero-shot fake news detection and demonstrates how combining individual models in ensembles can improvee performance. The results provide value to practical guidelines on the implementation of lightweight, misinformation detection systems that are domain-independent in real-life applications.

II. Literature review

Granik and Mesyura (2017) [3] have suggested the use of the Naive Bayes classifier to identify fake news on Facebook. They also trained their model on in-depth posts on different political pages to obtain a level of approximately 74 percent accuracy. But only 4.9 per cent of fake news was present in the dataset and this is what created the problem of class imbalance hence the false instances have become more difficult to classify. Although the proposed method utilized basic operations involving content-based features, the research set an initial benchmark of fake news identification and demonstrated that the conventional machine learning approaches do not work well with the inbalanced data [3].

Ahmed et al. (2017) [4] examined n-gram features using several classifiers, SVM, Logistic Regression, and kNN. With the 5-fold cross-validation, they discovered that the best result was by using linear SVM with unigram features, which had the highest performance- roughly 92 percent accuracy. They have revealed that simple lexical features combined with strong classifiers can be powerful and that this represents a big barometer to the classical supervised methods of fake news detection [4].

ZS-FND is a zero-shot fake news detection model proposed by Baashirah et al. (2024) [5], and is constructed on the basis of BERT-based embedding. Based on raw fake news datasets, ZS-FND performed at a state of the art-level 98.39%, 97.33%, and 96.49% in terms of accuracy, precision, and F1-score, respectively, without any parametrization steps. The MAE was also very low i.e. 0.0160, which shows that the model is very close to ground truth. The paper demonstrated the effectiveness of transformer-based models in identifying fake news when they do not have labeled data at all [5].

L-Defense is an explainable framework suggested by Wang et al. (2024) [6] to detect fake news with the help of LLMs. its integrates three modules namely evidence extraction, prompt-based reasoning and defense-based inference. It consists of a system that provides human-readable justifications based on information about the evidence provided in support and against it. L-Defense was evaluated on RAWFC and LIAR-RAW and reached the state-of-the-art accuracy and had better interpretability. The new novel metric called the discrepancy proved that fewer serious errors were confirmed. This strategy promotes effectiveness and openness in identifying fake news [6].

III. Methodology

Our method employs a multi-stage pipeline to detect fake news using a zero-shot approach. The pipeline processes a labeled news dataset through summarization, chunking, classification, ensemble aggregation, and evaluation steps. Initially, all news articles are summarized to shorten them and extract relevant details. The summarized text is split into smaller chunks, and each chunk is classified independently using zero-shot pre-trained language models. Chunk-level predictions are aggregated to determine the final label of the article. This system does not require task-specific fine-tuning for fake news classification.

A. Dataset

We used the Fake News Detection dataset from Kaggle, licensed under CC BY-4.0. It contains two CSV files: `fake.csv` and `true.csv`, including the title, full text, subject, and date of articles covering politics, health, and current events from 2015 to 2018. The `text` column was used for classification, while other columns were preserved for additional analysis.

B. Data Summarization

Each article was summarized using the PyTextRank algorithm with the SpaCy `en_core_web_lg` model. This extractive summarization method selects the most significant phrases and sentences from the original text to reduce length and remove irrelevant content. Summarized texts were saved in JSON files, with progress checkpoints every 2000 rows to allow resumption in case of interruptions.

C. Data Preprocessing

Summarized texts were cleaned by removing special characters and punctuation using regular expressions, converting all text to lowercase, and trimming leading and trailing spaces. Missing or empty entries were identified and dropped using Pandas' NA handling. To address class imbalance, random downsampling was applied to the majority class using Scikit-learn's `resample()` function, resulting in a balanced dataset suitable for classification.

D. Zero-Shot Classification Models

We employed four state-of-the-art zero-shot classification models to distinguish real from fake news without task-specific training. Since all models have a maximum input token length

TABLE I
Comparison of Approaches for Fake News Detection

Paper	Method	Key Features	Limitations
Granik et al. (2017) [3]	Naive Bayes Classifier	Simple real-time system, tested on social data	Skewed data (4.9% fake), low fake news accuracy
Ahmed et al. (2017) [4]	N-gram + ML (LSVM best)	6 classifiers, 5-fold CV, strong lexical baseline	Lacks deep semantics, dataset unclear
Baashirah (2024) [5]	Zero-Shot BERT (ZS-FND)	No task-specific training, low MAE (0.016), high F1	Dataset not disclosed, limited generalization
Wang et al. (2024) [6]	L-Defense (LLM-based)	LLM + prompt reasoning, interpretable, low Discrepancy	Accuracy not reported, high computation

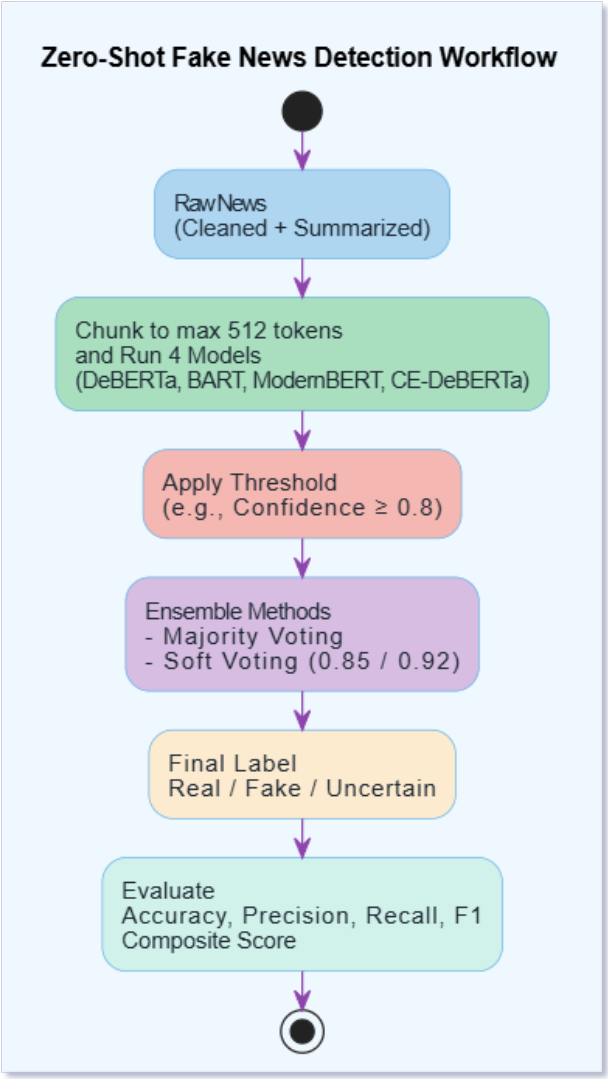

Fig. 1. Zero-Shot Fake News Detection Workflow

of 512, long articles were chunked accordingly. Each chunk was classified independently with "real" and "fake" candidate labels. The highest confidence score across chunks determined the final article label if it exceeded a threshold of 0.8; otherwise, the article was marked as uncertain. The models used are:

- **MoritzLaurer/DeBERTa-v3-large-mnli-fever-anli-ling-wanli:** Transformer-based model with a 512-token limit, classifying each chunk independently and aggregating max confidence scores.
- **facebook/bart-large-mnli:** Sequence-to-sequence transformer with the same token constraints, processing chunks similarly.
- **MoritzLaurer/ModernBERT-large-zeroshot-v2.0:** Zero-shot model designed for classification, employing chunk-wise max scoring.
- **cross-encoder/nli-deberta-v3-base:** Cross-encoder model that classifies chunks independently, selecting the highest confidence label above threshold.

Classification results were periodically saved to enable progress tracking and resumption.

E. Ensemble Learning

To improve classification robustness, we combined individual model predictions using two ensemble methods:

- **Majority Voting:** For each article, the four models vote on "real" or "fake," excluding uncertain predictions. The label with the majority of votes is assigned as the final classification. Articles with no majority were excluded from evaluation.
- **Soft Voting:** Sum of confidences between real and fake were added up across models. In case the better score was above a certain level (0.85 or 0.92), that label was selected, otherwise, the outcome was indicated uncertain, this way that allows having more accurate predictions.

F. Evaluation

We evaluated both individual models and ensemble methods using standard metrics: accuracy, precision, recall, F1-score and overall Performance Score. Predictions marked uncertain were excluded to focus on confident classifications. Confusion matrices were generated to analyze error patterns, showing true positives, true negatives, false positives, and false negatives for each approach.

For visual comparison, we used bar charts, radar charts, and heatmaps to present the performance metrics. Ensemble methods outperformed individual models, with majority voting improving classification consistency through consensus, and soft voting further enhancing accuracy by incorporating confidence scores.

IV. Results and Analysis

This part entails the experimental outcomes associated with the final run of the identified zero-shot classification models operating on the fake news detection dataset. Common evaluation metrics are used to measure performance, including *accuracy*, *precision*, *recall*, *F1-score*, and *uncertainty percentage*. The performances of individual models are compared, highlighting their strengths and weaknesses. Furthermore, an overview of ensemble methods is provided. Tables are incorporated to support the interpretation of results and to better illustrate performance differences among the models.

A. Dataset Overview

A preprocessing was performed in several steps to control the quality of data and class imbalance: removal of noise and downsampling. This guaranteed an adequate data set, a lower level of bias, and enhanced the integrity and fairness of the identification of model performance.

TABLE II
DATASET PREPROCESSING STAGES AND DISTRIBUTION

Stage	Fake	Real	Total Rows	Notes
Before Cleaning	23,481	21,417	44,898	Original dataset before any filtering or edits
After Cleaning	22,851	21,416	44,267	Removed noisy, empty, or duplicate entries
Before Downsampling	22,851	21,416	44,267	Cleaned data, but still imbalanced
After Downsampling	21,416	21,416	42,832	Balanced dataset with equal fake/real instances

B. Individual Model Performance

The current discussion addresses the issue of reliability of model categorization based on the evaluation of confusion matrices, confidence-accuracy relationships, and confidence-scores distributions. These analyses explain why model confidence with predictions occurs with high frequency and why they rather avoid delivering a judgment phrased with uncertainty, and such capabilities are required to perform tasks like fake news detection.

1) MoritzLaurer/DeBERTa-v3-large: The histogram in Figure.2 shows that the most confidence scores of DeBERTa-v3-large falls within 0.8 to 0.9, which denotes a great confidence in what the model predicts. This kind of distribution implies that the system might be useful in the filtering of uncertainty.However, caution should be exercised regarding how well these confidence levels reflect actual prediction accuracy.

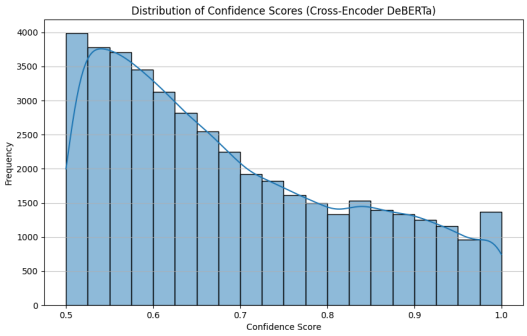

Fig. 2. Confidence score distribution of DeBERTa-v3-large predictions

The bias in the model is significant in the sense of recognizing real content with a high level of confidence and shows more difficulty as far as recognizing the fabricated pieces is concerned. Since the confusion matrix explicitly discards unclassified or uncertain cases, then it necessarily presents values of interest only over predictions made confidently. Accordingly, the results demonstrate the functionality of the system against strict reliability requirements. The high false-negative rate indicates that although it has clear calibration methods in place, the model still allows some fake news to continue given a high-confidence score.

2) facebook/bart-large-mnli: The histogram in Figure.4 shows that the confidence score in BART model are also clustered around 0.9, which means that high predictions were prevalent. This kind of focus is beneficial to filtering and to the rapid selection of probable hypotheses; but it also increases the dangers of over-confidence, and reflects the need of calibration

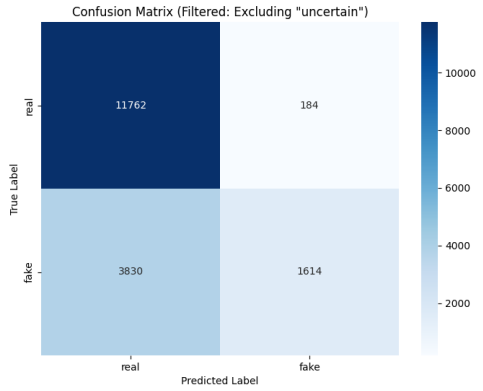

Fig. 3. Confusion matrix of DeBERTa-v3-large predictions with confidence threshold 0.8

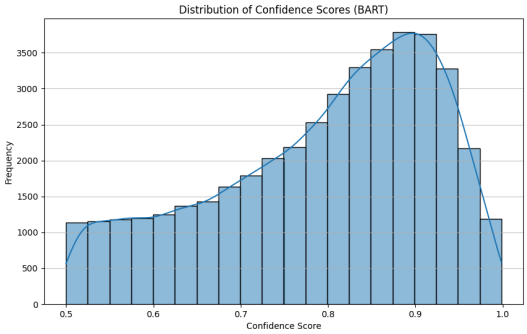

Fig. 4. Confidence score distribution of BART-large-MNLI predictions

to define accurately the limits either of uncertain, or of firm inferences.

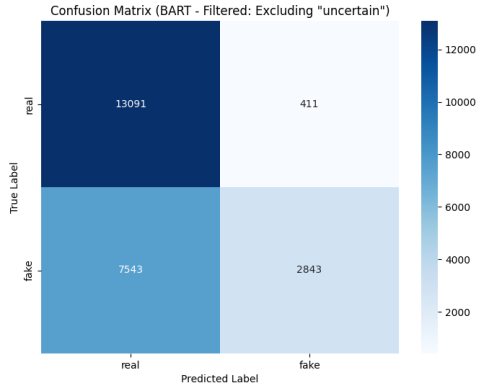

Fig. 5. Confusion matrix of BART-large-MNLI predictions (confident cases only)

The BART model shows a strong bias to categorizing of true reporting as a result of which it achieves high values of sensitivity concerning it. On the other hand, the model will display a low level of effectiveness in rejecting generated material when

prediction reliability is strong given the high percentage of false negatives. Since the matrix under consideration includes only high-confidence predictions, such findings reveal that the model is miscalibrated; a large proportion of classified fake news is assigned with unconfident predictions.

3) MoritzLaurer/ModernBERT-large-zeroshot-v2.0: The confidence histogram in Figure. 6 illustrates that ModernBERT frequently produces predictions with high certainty, peaking sharply at the 1.0 mark. While this assertiveness can be advantageous in selective prediction scenarios, it also raises concerns about overconfidence. Over-saturation in high certainty levels might lead to an overestimation of predictive accuracy, especially if not supported by empirical correctness. Therefore, additional calibration testing is essential to verify that these confident scores genuinely correspond to accurate predictions, particularly in domains such as misinformation detection where errors can be critical. The model is biased

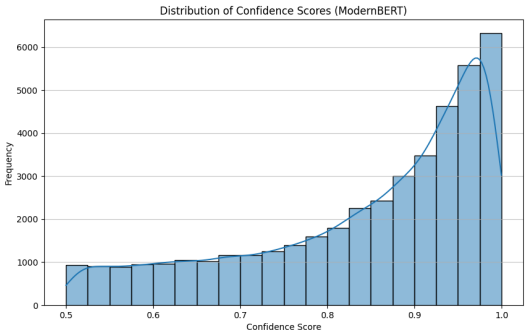

Fig. 6. Distribution of confidence scores for ModernBERT predictions

toward identifying real content with a high likelihood but hesitates to label news as fake. A large number of false negatives were observed, indicating potential issues such as class imbalance, overlapping features, or skewed confidence calibration. The confusion matrix in Fig. 7 reflects only confident predictions, as uncertain ones were filtered out. Thus, the matrix represents the most reliable behavior the model can exhibit. Despite this, many fake news instances are still misclassified, suggesting areas for improvement.

4) cross-encoder/nli-deberta-v3-base: The histogram in Fig. 8 shows the distribution of confidence scores for this model. Most scores are concentrated in the 0.5–0.6 range, indicating a cautious prediction tendency. While this conservativeness may help prevent overconfident misclassifications.

The confusion matrix in Fig. 9 visualizes the classification patterns for confident predictions. Since predictions below the 0.8 confidence threshold were filtered out, this matrix provides insight into the model's behavior under certainty. It suggests that while the model performs well overall, there remains a tendency to falsely flag true information, a consideration important for practical deployment.

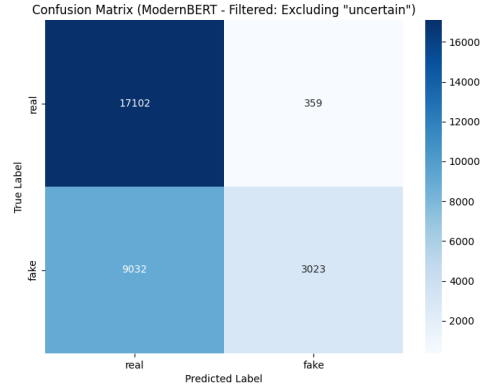

Fig. 7. Confusion matrix of confident predictions by ModernBERT

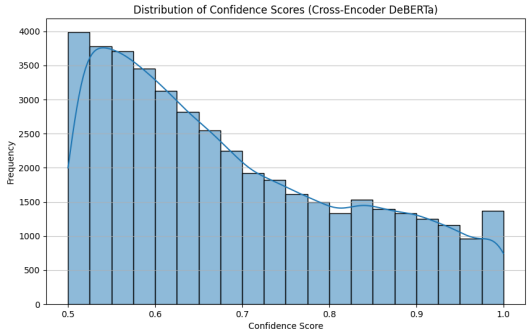

Fig. 8. Confusion matrix of Cross-Encoder DeBERTa (confident predictions)

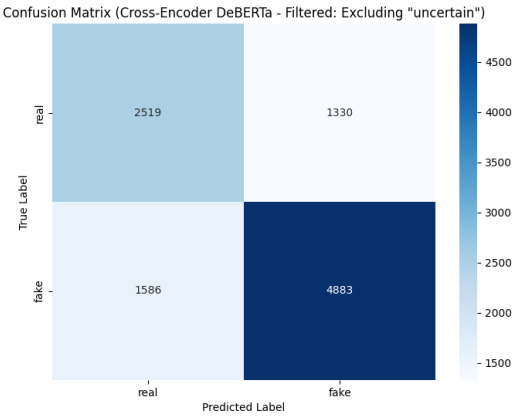

Fig. 9. Confusion matrix of Cross-Encoder DeBERTa (confident predictions)

TABLE III
PERFORMANCE COMPARISON OF ZERO-SHOT MODELS (CONFIDENCE THRESHOLD = 0.8)

Model	Accuracy	Precision	Recall	F1-Score	Uncertain
DeBERTa	0.7692	0.7992	0.7692	0.7264	0.5940
BART	0.6670	0.7385	0.6670	0.6148	0.4423
ModernBERT	0.6818	0.7522	0.6818	0.6241	0.3109
CE-DeBERTa	0.7174	0.7217	0.7174	0.7191	0.7591

C. Zero-Shot Model Comparative Analysis

Figure. 10 uses a radar plot to map model capabilities across five evaluation axes. Both CE-DeBERTa and DeBERTa span wider areas, highlighting strong predictive potential. However, CE-DeBERTa's elevated uncertainty is visible, reinforcing the notion that performance often comes at the cost of decisiveness.

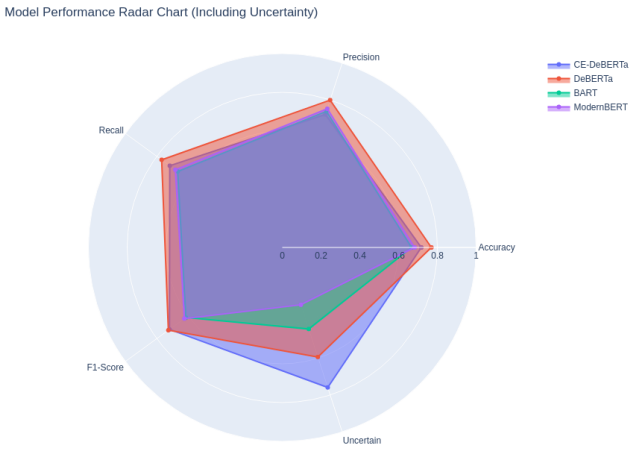

Fig. 10. Radar Plot: Transformer Model Performance and Uncertainty

D. Performance of Ensemble Model

TABLE IV
ENSEMBLE METHOD PERFORMANCE COMPARISON

Metric	Majority Voting	Soft Voting (0.85)	Soft Voting (0.92)
Accuracy	0.6697	0.6783	0.7023
Precision	0.7336	0.7355	0.7540
Recall	0.6697	0.6783	0.7023
F1-Score	0.6290	0.6388	0.6615
Uncertainty	0.1525	0.2080	0.3199
Certainty (1 Unc.)	0.8425	0.7920	0.6801
Predictions Evaluated	36,087	33,923	29,132
Performance Score	0.7089	0.7046	0.7000

To enhance the robustness of predictions and mitigate individual model biases Ensemble Method is applied here. A comparison of three methods of ensembling, Majority Voting and Soft Voting with the thresholds of 0.85, 0.92 is provided in Table IV. When Soft Voting threshold is raised, accuracy, precision, recall and F1-score measurements are also higher which means that stricter confidence requirements imply better classification. At the same time, it comes at an expense of higher uncertainty, i.e. more predictions will be marked as uncertain and fewer of them will be definitely examined. This trade-off is indicated in a certainty metric that will fall to 0.92 in the case of Majority Voting of 84.25 percent to Soft Voting of 68.01 percent. The average of this (and some other measures), weighted as the performance score, establishes that Majority Voting has the best balance overall, of accuracy and confidence. Therefore, Majority Voting offers the most confident and reliable predictions and, hence, this method is recommended to detect fake news where it is critical to have trustworthy findings.

E. Threshold and Uncertainty Effect

A heatmap in figure 11 is introduced to compare in terms of Accuracy, Precision, Recall, F1-score, and Uncertainty three ensemble strategies that include Majority Vote, Soft Vote (Threshold=0.85) and Soft Vote (Threshold=0.92). The best predictive result is exhibited by the approach Soft Vote with Threshold=0.92: Accuracy 0.702, Precision 0.754, Recall 0.702, F1-score 0.661. This is however associated with the greatest Uncertainty (0.320) which depicts that a more rigid cut is what augments dispersion of prediction model confidence. Demanding more consensus among the base predictions, there will be less confident allocation of examples, such that it returns a high performance on a few instances, but a higher Uncertainty overall. Majority Vote may obtain a lesser Uncertainty (0.157) with a loss of performance overall. Moderate values and Uncertainty (0.208) are presented by the intermediate threshold (0.85) and the associated trade-offzone. In turn, raising the threshold intensifies prediction Confidence, but raises Reliability on high-Confidence samples at the cost of high U to low generalizability. The best approach will depend on whether an application has a greater emphasis regarding precision or confidence calibration.

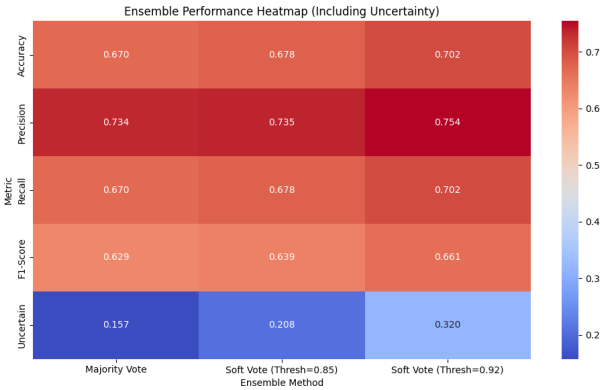

Fig. 11. Heatmap Comparison of Ensemble Voting Strategies Across Evaluation Metrics

F. Limitations

Despite improved reliability through ensemble approaches, several limitations remain:

- The models are built on pretrained architectures, which may encode linguistic or domain-specific biases.
- Confidence scores, while useful, are not always reliable indicators of correctness and can suffer from miscalibration.
- The dataset used may not adequately represent the diversity of real-world misinformation scenarios.
- This study employed fixed thresholds; dynamic or adaptive thresholding strategies were not explored.

Future work should consider adaptive calibration techniques, multi-lingual robustness, and domain-specific fine-tuning to

build more resilient and accurate misinformation detection systems.

V. Conclusion and future works

This work shows that ensemble learning by setting confidence thresholds and uncertainty analysis can play a very crucial role in detecting misinformation that uses transformer models. Through the comparison of voting approaches and the thresholds we determined some of the main trade-offs between performance and reliability. A Majority Voting enables more predictions at the expense of scarcity, and Soft Voting with increased thresholds will increase precision and F1-score at the cost of certainty. The visualizations, which were used to substantiate these dynamics, included the confidence histograms and the radar plots. Adaptive thresholds conditioned on the attribute of inputs, uncertainty-equivalent loss functions, and task-specific fine-tuning are the subjects of future research. Improving the generalizability using training datasets composed of multiple languages and adding time or user metadata can also help increase robustness and realistic application of fake news detection.

VI. Discussion

The findings outline the significance of the certainty of models in the ensemble technique of fake news detection. Although the performance of the deeper models (such as CE-DeBERTa) is high raw, they have higher uncertainty; hence it seems that careful application should be used. Ensemble methods can be used to slightly compensate this by adjusting the predictions to the degree of confidence reached, providing a balance between boldness and the reluctance to jump to conclusions. This trade-off is vital in the real-world scenarios, as it is possible to send high-confidence results to human moderators and yet to provide additional attention to uncertain examples. Measurements of uncertainty do not only enhance the readability, but also aid in strategic decisions, particularly in areas of social sensitivity such as misleading detection, where a failure to detect the latter can have grave social and credibility costs.

References

[1] A. Gupta, H. Lamba, and P. Kumaraguru, "Emerging threats on twitter: A survey of fake news detection," *Computers & Security*, vol. 97, p. 101 947, 2020.

[2] K. Shu, D. Mahudeswaran, S. Wang, and H. Liu, *FakeNewsNet: A data repository with news content, social context and spatiotemporal information for fake news research*, https://github.com/KaiDMML/FakeNewsNet, Accessed: 2025-05-13, 2018.

[3] M. Granik and V. Mesyura, "Fake news detection using naive bayes classifier," May 2017, pp. 900–903. DOI: 10.1109/UKRCON.2017.8100379.

[4] H. Ahmed, I. Traore, and S. Saad, "Detection of online fake news using n-gram analysis and machine learning techniques," in *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments: First International Conference, ISDDC 2017, Vancouver, BC, Canada, October 26-28, 2017, Proceedings 1*, Springer, 2017, pp. 127–138.

[5] R. Baashirah, "Zero-shot automated detection of fake news: An innovative approach (zs-fnd)," *IEEE Access*, vol. 12, pp. 182 828–182 840, 2024. DOI: 10.1109/ACCESS.2024.3462151.

[6] B. Wang, J. Ma, H. Lin, *et al.*, "Explainable fake news detection with large language model via defense among competing wisdom," in *Proceedings of the ACM Web Conference 2024*, 2024, pp. 2452–2463.