AI-Powered Fact Verification: Analyzing Public Consensus Through Machine Learning

Aakash Mor
University of the Arts London
aakashmor@gmail.com

***Abstract*—In the face of rising misinformation across social media platforms, there is a growing need for scalable and automated systems that can verify factual accuracy. This study introduces a machine learning framework that leverages a transformer-based model to evaluate how well social media posts align with factual statements. The framework is tested across various classification strategies, including sentiment analysis, consensus scoring, and both binary and multi-class categorization. Among these approaches, binary classification demonstrated the strongest performance in detecting factual consistency. While sentiment-based methods offered limited insight, they were not sufficient to determine truthfulness on their own. The results highlight the potential of interpretable, high-accuracy models in supporting large-scale efforts to counter misinformation.**

I. Introduction

The proliferation of digital communication channels, particularly social media platforms, has intensified the spread of misinformation and disinformation. This surge poses a critical threat to public discourse and institutional trust. In response, there is a pressing need for intelligent systems that can assess and verify the factual integrity of online statements. The challenge is particularly acute on platforms like Twitter, where information is disseminated rapidly with limited context.

This study approaches the problem through the lens of computational linguistics and machine learning, employing the DistilBERT transformer model [1] to evaluate the degree of factual alignment between social media content and corresponding claims. The focus lies in predicting the extent of semantic congruence—or "accord"—between tweet text and paired declarative statements.

Using a dataset of tweet-claim pairs labeled with human-assessed agreement levels, the research explores preprocessing strategies and their influence on classification performance. The key contributions include a comparative analysis of modeling techniques, an investigation into label complexity and input representation, and suggestions for refining dataset design to improve real-world applicability in automated fact-checking tasks.

II. Related Work

Artificial Intelligence continues to find applications across diverse fields—from digital security to health diagnostics and information verification. Several researchers have contributed to advancing these domains through domain-specific adaptations of machine learning techniques.

Saimbhi [2], [3] has worked on software security and digital media validation. His approach to vulnerability detection involves integrating various code representations—abstract syntax trees, control flow graphs, and dependency models—into a unified graph structure, optimized using convolutional neural networks. In a separate study on image authenticity, he uses Discrete Cosine Transform (DCT) techniques along with machine learning classifiers to distinguish between real and manipulated UHD visuals.

Desai [4], [5] focuses on practical implementations of AI. His work includes developing Progressive Web Applications (PWAs) for inventory tracking using geolocation and barcode scanning features. He also investigates active learning strategies to optimize text classification processes and applies Random Forest algorithms to identify pricing determinants in consumer retail, using IKEA products as a case study.

Katta [6]–[8] expands the reach of AI into medical diagnostics, financial forecasting, and decentralized learning. Her work on early-stage lung cancer detection utilizes transfer learning and deep neural networks to improve accuracy in CT scan analysis. In federated learning, she proposes AsyncHierFed, a decentralized model that balances performance across distributed systems. Additionally, she offers predictive models for analyzing high-frequency cryptocurrency trading using WebSocket data.

Collectively, these efforts demonstrate the versatility of machine learning across both theoretical and applied contexts, underscoring its relevance to misinformation analysis.

III. TruthSeeker Dataset

The TruthSeeker dataset is a curated resource designed to support the development of models for misinformation detection, particularly on Twitter. It consists of tweet-claim pairs, each labeled based on the perceived relationship between the tweet and a declarative statement. This structure allows for the training of classification models that assess the factual alignment of short-form content.

A. Terminology Reassessment

Although the dataset offers strong potential for misinformation research, a closer examination of its terminology reveals areas for improvement.

Binary View of "Truth": Labeling content as simply "true" or "false" reduces the complexity of factual discourse to rigid

categories. Digital communication often involves nuance, context, and partial truths. While binary classification simplifies computational tasks, it may not reflect the multifaceted nature of real-world information.

Limitations of "Ground Truth": The dataset frequently refers to "ground truth" as an absolute standard based on expert fact-checking. However, validation often depends on interpretive judgments that can vary between organizations and evaluators [**?**]. A probabilistic or confidence-based framing may better reflect the inherent uncertainties in truth assessment.

Ambiguity in "Agreement": The use of the label "agreement" to describe tweet-claim relationships may conflate subjective agreement with objective accuracy. Since annotations are crowd-sourced, individual interpretations, political biases, and contextual gaps may affect labeling consistency. Replacing "agreement" with clearer terms such as "alignment rating" or "evidence-based match" would improve the dataset's semantic clarity.

Addressing these concerns will strengthen the dataset's conceptual integrity and enhance its usefulness for building more accurate and reliable misinformation detection systems.

B. Data Preparation Summary

The dataset construction process began by collecting 700 verified true and 700 verified false articles from PolitiFact. For each article, two to five manually selected keywords were used to extract relevant tweets using the Twitter API. Although several automated keyword extraction tools—including RAKE, PKE, and YAKE—were tested, they either returned irrelevant results or insufficient matches. Manual keyword curation proved more effective, ultimately yielding a dataset of approximately 186,000 tweets, with an average of 133 tweets associated with each article. This comprehensive corpus provides a solid foundation for evaluating semantic alignment between user-generated content and factual claims.

C. Crowdsourced Annotation Strategy

To annotate tweet-claim pairs, Amazon Mechanical Turk (MTurk) was used, restricting the task to "Master" workers to ensure consistency and accuracy. Annotators were instructed to assign one of four stances—*True*, *False*, *Mostly True*, or *Mostly False*—based on each tweet's relation to the associated statement. To reduce subjectivity, each item was reviewed by three different workers, and final labels were derived through majority voting.

D. Data Cleaning and Label Structuring

Entries marked as "NO MAJORITY" or labeled "Unrelated" were removed to maintain quality. A "ground truth" column was introduced to reflect consensus judgments, serving as the primary target label for model training. After filtering, the dataset included roughly 150,000 annotated tweets aligned with 1,400 verified claims, with a balanced label distribution.

TABLE I
FOUR-CLASS LABEL MAPPING

Statement Type (T/F)	Crowd Judgment
T	Agreement
T	Contradiction
T	Partial Agreement
T	Partial Disagreement
F	Agreement
F	Contradiction
F	Partial Agreement
F	Partial Disagreement

TABLE II
BINARY LABEL MAPPING

Statement Type (T/F)	Crowd Judgment
T	Agreement
T	Contradiction
F	Agreement
F	Contradiction

E. Dataset Enhancements

Initially, training combined both the original statement classification and majority-vote labels. Later, it was found that using only the annotator consensus improved model performance, leading to a revised labeling strategy. To prevent data leakage, claim-level separation was enforced across training, validation, and testing splits.

Further improvements included generating a composite input feature combining tweet text, claim, and consensus label, allowing the model to better learn the semantic relationships involved. The dataset was also reformatted into a 'Dataset' object compatible with the `datasets` library, supporting seamless integration with transformer-based pipelines.

IV. EVALUATION OVERVIEW

The evaluation aimed to replicate the baseline framework from the original TruthSeeker study, followed by modifications to test performance gains. Due to submission deadlines, only selected refinements were completed. Below is a summary of the experiments and their outcomes.

A. Experiment 1: Tweet-Only Agreement Classification

This baseline experiment predicted "agreement" using only tweet text, omitting claims for contextual grounding. Labels were derived from the consensus annotation (see Tables I and II). A DistilBERT model [1] was fine-tuned using accuracy as the primary metric.

a) Findings:: Performance was low (around 30% accuracy), likely due to insufficient context. Tweets alone, being brief and ambiguous, lacked the depth needed to infer factual stance accurately. This setup will be revisited in future work under improved preprocessing conditions.

B. Experiment 2: Multi-Class (Four-Label) Classification

The model was trained to classify tweet-claim pairs into four categories: *True*, *Mostly True*, *False*, and *Mostly False*. Input was reformatted to combine the tweet, statement, and consensus label into one sequence.

a) Training Details:: Hyperparameter tuning began with a learning rate of 1×10^{-5}, later reduced to 1×10^{-8} for better convergence. Despite the richer input, model improvement stagnated after the first epoch.

b) Findings:: Achieving 49% accuracy, the model struggled with fine distinctions, particularly between "True" and "Mostly True" or "False" and "Mostly False." Subjectivity in crowd labels likely contributed to classification ambiguity. Future enhancements may include label smoothing or more robust architectures.

C. Experiment 3: Binary Classification

Simplifying the task to a binary setup, the model classified tweets as either "True" or "False," based on the consensus labels.

a) Approach:: The binary cross-entropy loss function replaced categorical cross-entropy, while all preprocessing and input formatting remained consistent with the multi-class experiment.

b) Results:: The binary model performed best, achieving an accuracy of 96%. Early convergence indicated the model quickly learned strong decision boundaries. This result highlights the trade-off between label granularity and model performance.

D. Experiment 4: Sentiment-Based Prediction

This experiment tested whether tweet sentiment could infer agreement with claims. Two methods were used: VADER [9] for lexicon-based scoring, and DistilBERT for fine-tuned sentiment classification.

a) Outcome:: VADER produced general emotional orientation but lacked alignment with factual correctness. DistilBERT modeled sentiment well but failed to correlate consistently with truthfulness. These findings suggest that sentiment alone is an unreliable predictor for fact verification.

TABLE III
PERFORMANCE METRICS FOR SENTIMENT-BASED CLASSIFICATION

Label	Precision	Recall	F1 Score	Count
False	0.74	0.61	0.67	41
True	0.69	0.80	0.74	45
Overall Accuracy	0.71 (Total Instances: 86)			
Macro Average	0.71	0.70	0.70	86
Weighted Average	0.71	0.71	0.71	86

b) Findings: The sentiment-based classifier demonstrated moderate overall performance. Precision was slightly higher for the "False" class (0.74) than for "True" (0.69), indicating a slight preference for accurately identifying negative sentiment. Conversely, recall was stronger for the "True" class (0.80), suggesting a higher sensitivity toward detecting positive sentiment. The F1 scores followed a similar trend, with 0.74 for "True" and 0.67 for "False".

Although a 71% accuracy suggests some utility, the findings underscore the limitations of sentiment analysis as a standalone method for factual assessment. Emotional tone often fails to reliably signal truthfulness or alignment with declarative claims. For more robust performance, sentiment cues must be combined with contextual and content-specific signals such as claim semantics, evidence traces, or linguistic patterns [10], [11].

E. Comparative Evaluation Summary

Among the evaluated methods, binary classification proved most effective. Sentiment-based approaches offered moderate results but failed to capture the deeper factual structure required for accurate verification. Results suggest that integrating context and structure is crucial for reliable classification.

V. Model Limitations and Discussion

The study revealed several limitations:

- **Annotation Bias:** Crowd-sourced labeling, while scalable, introduced subjective noise. Majority judgments may inadvertently reflect prevalent biases or misinterpretation.
- **Overfitting in Multi-Class Models:** In the four-class setup, overfitting was evident beyond initial epochs, highlighting the need for improved generalization strategies.
- **Sentiment Misalignment:** Relying on emotional polarity as a proxy for factual correctness led to inconsistencies. Sentiment alone lacks the depth to capture nuanced claim alignment.
- **Limited Input Representation:** Early experiments that excluded claim context showed poor results, reinforcing the importance of comprehensive input features.

VI. Implications

This work has broader implications for the deployment and ethical considerations of misinformation detection systems:

- **Bias Propagation:** Overreliance on consensus labels may embed sociocultural or ideological biases into automated decisions.
- **Ethical Concerns:** In sensitive domains like health or politics, mislabeling may reinforce misinformation or marginalize dissenting voices.
- **Impact on Discourse:** Automated moderation may suppress minority perspectives, leading to echo chambers and a lack of viewpoint diversity.
- **Adversarial Risk:** Models may be vulnerable to manipulative inputs designed to evade or deceive automated detection.
- **Need for Transparency:** As such systems scale, ensuring model interpretability, fairness, and accountability is paramount.

VII. Future Work

Several directions can enhance this framework:

- **Enhanced Sentiment Modeling:** Fine-tune RoBERTa or domain-adapted BERT variants for sentiment detection specific to political or health-related discourse.
- **Adversarial Robustness:** Evaluate performance against adversarial examples to improve model resilience and trustworthiness.

TABLE IV
EVALUATION SUMMARY OF DIFFERENT CLASSIFICATION STRATEGIES

Method	Accuracy	F1 Score	Observations
Binary Classification	96%	0.93	Delivered the highest accuracy. Clear label boundaries facilitated efficient learning and consistent results.
Four-Class Categorization	49%	0.47	Reduced accuracy due to subjective overlaps between categories like "Mostly True" and "True."
Tweet-Only Consensus	30%	0.29	Performed poorly as it lacked contextual grounding from the source claims.
Sentiment-Based Analysis	71%	0.74 (True) / 0.67 (False)	Useful for gauging emotional tone but insufficient as a factual alignment mechanism.

- **Multimodal Features:** Integrate metadata, social engagement patterns, and user behavior for richer contextual grounding.
- **Gradient Agreement Scores:** Move beyond categorical outputs to predict degrees of alignment, useful for opinion mining or policy monitoring.
- **Cross-Domain Deployment:** Apply models in journalism, scientific consensus tracking, or crisis response with appropriate tuning.

Planned Contributions for Next Iteration

To strengthen the current framework, future iterations aim to:

- Merge sentiment analysis with contextual modeling for hybrid detection systems.
- Incorporate social network signals and content propagation pathways.
- Conduct real-time testing under adversarial constraints.
- Tailor the system for domain-specific fact-checking, such as health communication or public science reporting.

VIII. CONCLUSION

This study reconstructed and extended the TruthSeeker fact-verification system using DistilBERT and VADER models. It evaluated the role of sentiment as a feature for factual alignment and assessed classification strategies ranging from binary to multi-class models.

Results confirmed that sentiment signals offer some predictive value but are insufficient alone for high-stakes veracity tasks. Binary classification emerged as the most effective approach, demonstrating the need for simplicity, clear labels, and contextual information.

Future work will focus on refining label structures, enhancing robustness, and exploring hybrid models that integrate emotion, logic, and evidence to support more trustworthy AI-driven content verification systems.

REFERENCES

[1] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[2] A. S. Saimbhi, "Enhancing software vulnerability detection using code property graphs and convolutional neural networks," in *2025 International Conference on Computational, Communication and Information Technology (ICCCIT)*, 2025, pp. 435–440.

[3] ——, "Distinguishing true and fake ultra-high definition images using relative dct analysis and machine learning," in *2025 International Conference on Emerging Systems and Intelligent Computing (ESIC)*, 2025, pp. 172–178.

[4] A. Desai, "Enhancing inventory management with progressive web applications (pwas): A scalable solution for small and large enterprises," in *TechRxiv*, 2025.

[5] ——, "Unveiling the drivers of ikea product pricing: A random forest analysis," in *TechRxiv*, 2025.

[6] K. Katta, "Deep learning for early lung cancer detection from ct scans: A data science bowl approach," in *2024 IEEE International Conference on Future Machine Learning and Data Science (FMLDS)*, 2024, pp. 308–314.

[7] ——, "Asynchronous hierarchical federated learning: Enhancing efficiency in distributed learning systems," in *2024 5th International Conference on Computers and Artificial Intelligence Technology (CAIT)*, 2024, pp. 462–469.

[8] ——, "Forecasting returns for high-frequency cryptocurrency websocket data," in *2024 10th International Conference on Computer and Communications (ICCC)*, 2024, pp. 377–386.

[9] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 8, no. 1. Association for the Advancement of Artificial Intelligence, 2014, pp. 216–225.

[10] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," *LREC Workshop*, 2010. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2010/workshops/W3.pdf

[11] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1-2, pp. 1–135, 2008.