Project 1 Deliverable 2: Detailed Technique Report

The algorithm

Overview

The system ingests a URL or text, extracts structured features, computes a rule-based baseline score, optionally runs machine learning models to produce a ML score, and returns a final hybrid score with an explanation.

Feature extraction

The algorithm begins by extracting credibility-related features and returning a score and details for each. It accepts a url or or text as input. If input is a url, the algorithm uses *requests* to retrieve and *BeautifulSoup* to extract text from the webpage. If the url cannot be fetched or there is no text content, an error is returned.

The first feature evaluated is source authority. The algorithm gives a source authority score of 9 to urls that have “gov” or “edu” in the suffix, and 7 to urls that have “org”. Everything else (“com”, “net”, etc.) gets a source authority score of 5. The score is out of 10, so in this case, 9 is a high score, i.e., highly credible, while 5 is neutral, i.e., not particularly credible or not credible.

The second feature looked at is publication quality based on word count and potential ads. Here, a neutral quality score of 5 is set to start. Two points are added if word count is greater than 800, and two points are subtracted if ad elements are greater than 10.

Next, citation patterns are evaluated. The number of links to external sources are counted and that number, if less than 10, becomes the citation score. If there are 10 or more links, the score is 10.

Finally, the content is scored on accuracy based on suspicious keywords. The number of times "miracle cure", "shocking secret", "click here", or "you won’t believe" occur in the text is counted. A score of 8 is set to start, and the number of suspicious keywords counted is deducted.

The four scores and the details that determine the scores are returned. For example:

{'scores': {'source_authority': 5, 'publication_quality': 5,

'citation_patterns': 10, 'content_accuracy': 8},

'details': {'domain': 'bbc.com', 'word_count': 2111,

'external_links_count': 138, 'suspicious_keywords_found': 0}}

Rule-based score

A rule-based score is computed given the four extracted features. The features are weighted 30, 25, 25, and 20 percent respectively, and a normalized score between 0 and 1, rounded to two decimal places, is returned.

Hybrid credibility assessor

The next part of the algorithm introduces the main function that combines a machine learning approach with the previously returned rule-based score, and returns a final score. It checks for an existing machine learning model. If one exists, an array is created consisting of the four features which is used as input to predict the machine learning score. The machine learning score and rule-based score are then weighted–60 and 40 percent respectively–to result in the final score. If there is no machine learning model available and thus no machine learning score, the final score is equal to just the rule-based score.

Build explanation

In addition to the final score, the output of the algorithm includes an explanation highlighting the features that were most impactful on the score. For example, if the source authority score is greater than or equal to 7, the string “source has strong domain authority" will be appended to the explanation; if the citation patterns score is greater than 5, the string "it provides external references" will be appended to the explanation, and so on. The final output is a JSON object that includes the score and explanation.

Demo

The final part of the algorithm runs the credibility assessor on three example inputs.

Parameters

- Rule weights can be adjusted to change the influence of each feature on the rule-based score.
- The classification probability cutoffs for the machine learning model can be tuned for balancing precision/recall.

- The weighting between rule-based and machine learning components can be optimized as well.

Maintenance and refinement

- As web content evolves, feature extraction can be expanded or adjusted in both the rule-based and machine learning components.
- Machine learning model should be retrained periodically as new labeled data becomes available.
- Rule sets should be reviewed and expanded if needed to reflect new credibility standards.
- Model calibration can be validated using benchmark datasets (e.g., LIAR, FakeNewsNet).
- Audit regularly for bias, ethical fairness, and explainability compliance.
- Integrate multimodal signals (text, images, social metadata).
- Enhance temporal adaptability to evolving credibility contexts.
- Develop richer evidence-retrieval and reasoning layers.
- Incorporate uncertainty estimation in the output for higher trustworthiness.

Existing models and techniques

Through academic research and industry implementations have come about widey-used practices for credibility assessment. These include rule-based/heuristics systems, feature-based machine learning, deep learning claim verification, social context and propagation models, and industry products that combine manual auditing with automated signals and tooling for fact-checkers. Hybrid & ensemble systems combine rule heuristics, ML classifiers, knowledge checks, and social signals into blended scores.

- Rule-based/heuristic systems involve scoring a source based on domain signals (e.g., TLD, domain age, WHOIS, publisher reputation), text heuristics (e.g., word counts, reading grade, sensational keywords, ad density), and link patterns (inbound/outbound links, citation counts).
- In feature-based machine learning, lexical, syntactic, metadata and network features are extracted and used for supervised classification (e.g., SVM, Logistic Regression, Random Forest). The LIAR dataset contains over 12,800 short political statements with labeled attributes, and is a benchmark for training models for fact-checking.
- Deep learning is used to create transformer pipelines for evidence retrieval and claim verification. Fine-tuned transformers such as BERT, RoBERTa, DeBERTa

are commonly used. FEVER is a largescale dataset consisting of altered Wikipedia sentences used to train models to determine the veracity of a claim by finding relevant evidence from Wikipedia documents.
- Social contex and propagation tracing models are used to assess credibility by observing how claims spread (e.g., retweet graphs, user credibility, propagation speed), and stance classification (i.e., how replies and comments endorse or deny a claim). PHEME is a dataset containing rumors and non-rumors posted during breaking news events, and has been used extensively by researchers to train and evaluate new rumor detection models.
- Industry products like NewsGuard and Google Fact Check tools use a mix of automated credibility signals and expert human review.

Gaps and limitations with current approaches

- Coverage vs. precision trade-off: fact-checking is precise but covers few claims. Automated systems cover more but suffer from lower precision and high false positives.
- Generalization across domains and time: model performance is inconsistent when trained on, say, political conten, and implemented on content of other subjects (e.g., health, sceince, entertainment). As time goes on, new terms, events, outlets, etc. can cause degradation of performance.
- Data scarcity and label bias: high-quality labeled datasets are hard and expensive to come by. There is a lack of large amounts of data that is good quality, complete, without noise and not biased.
- Explainability vs. performance: high-performing deep models are less interpretable. It is not always easy for users of these methods to understand the reasons for the results.
- Manipulation and adversarial actors: bad actors adapt (bots, coordinated sharing, mimicry of reputable sources). Models that rely on social or lexical cues can be gamed.
- Multimodality: many systems focus on text only and do not account for image and video content.
- Evidence retrieval & reasoning: claim verification requires retrieving correct evidence; evidence retrieval remains a bottleneck.
- Operational constraints: deployment is complicated due to real-time scoring, API rate limits, privacy concerns, and the need for human review.
- Calibration & uncertainty: many models output uncalibrated scores without explicit indication when the model is unsure.

- Ethical and fairness concerns: risk of censorship, bias in labels and disproportionate impact on certain publishers.

Rule-based/ML hybrid approach for addressing limitations

The rule-based approach utilizes predefined rules and provides an interpretable baseline for determining credibility. The machine learning approach uses a model to make predictions based on patterns learned from existing data. Each method if used alone may have shortcomings. For example, the rule-based method will give a neutral source authority score to a website like nytimes.com even though it is known to be highly reputable. Combining it with the machine learning method will make up for this flaw. Used alone, the machine learning method will not function well if the input data is sparse or unavailable. In those cases, combining it with the rule-based method is crucial. When the machine learning model fails, rule-based still provides results. The hybrid approach avoids these inevitable issues. It combines the strengths of each method resulting in a robust credibility assessment. It best addresses the gaps and limitations:

- Coverage vs. precision trade-off: while rules give better precision, machine learning gives better coverage.
- Generalization across domains and time: while rules provides stability, machine learning provides adaptability.
- Data scarcity and label bias: rules work fine with minimal data while machine learning learns from richer signals.
- Explainability vs. performance: rules are transparent and easily interpretable–results can be explained fully, while machine learning improves performance.
- Manipulation and adversarial actors: rules detect anomalies and machine learning identifies subtle manipulation.
- Multimodality: machine learning can process multimodal inputs including text, image, video and more.
- Evidence retrieval & reasoning: machine learning learns evidence relationships automatically reducing the bottleneck issue. It infers reasoning while rules validate evidence links.
- Operational constraints: rules are fast, lightweight and scalable. Machine learning scales well. Overall the hybrid approach scales effectively with interpretable outputs.
- Calibration & uncertainty: results of rules are deterministic with no uncertainty. Machine learning adds confidence.

- Ethical and fairness concerns: fairness can be upkept via transparency of rules, unbiased data and balanced sampling.

References

https://pmc.ncbi.nlm.nih.gov/articles/PMC8935100/?utm_source=chatgpt.com

https://arxiv.org/pdf/1812.00315

https://www.wired.com/story/gallup-poll-fake-news-ratings/?utm_source=chatgpt.com

https://www.newsguardtech.com/solutions/news-reliability-ratings/?utm_source=chatgpt.com