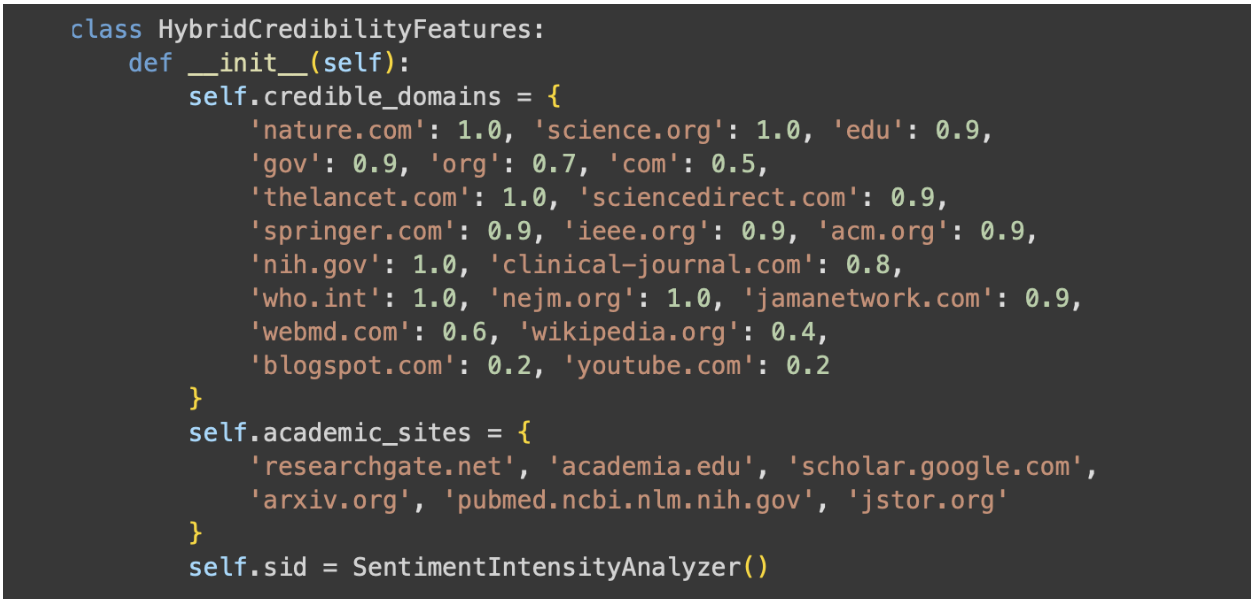
Raven Glenn
CS676: Algorithms for Data Science
Fall 2025

Project 1 Deliverable 2: Detailed Technical Report

For Project 1, we were given an objective to create an algorithm that helps one determine the credibility of a source given to them through a chatbot search. The algorithm which was derived adopts a hybrid approach that combines domain-based credibility evaluation, sentiment analysis, and supervised regression modeling to assign a credibility score on a scale from 0-100 to search results.

This proposed model integrates three major components: Hybrid Feature Engineering, Normalization and Model Training, and Prediction and Scoring. Hybrid feature engineering extracts credibility indicators from both URLs and linguistic content. Normalization and model training uses min-max scaling and linear regression for interpretable learning. Prediction and scoring produces rescaked credibility scores for new, unseen data. This pipeline emulates human-like reasoning; evaluating who is providing the information (source trustworthiness) and how it is presented (linguistic neutrality).

The system maintains a curated dictionary of known and trusted domains, each mapped to a numerical credibility weight. For a domain type that is considered to have high-credibility (ie. nature.com, science.org, nih.gov, who.int) they have been assigned scores between 0.9 and 1.0. Domains with moderate-credibility (ie. .edu, .org, clinical-journal.com) have been assigned scores between 0.5 and 0.8. Lastly, domains with low-credibility (ie. youtube.com, wikipedia.org) have been assigned scores below 0.5.



```
class HybridCredibilityFeatures:
    def __init__(self):
        self.credible_domains = {
            'nature.com': 1.0, 'science.org': 1.0, 'edu': 0.9,
            'gov': 0.9, 'org': 0.7, 'com': 0.5,
            'thelancet.com': 1.0, 'sciencedirect.com': 0.9,
            'springer.com': 0.9, 'ieee.org': 0.9, 'acm.org': 0.9,
            'nih.gov': 1.0, 'clinical-journal.com': 0.8,
            'who.int': 1.0, 'nejm.org': 1.0, 'jamanetwork.com': 0.9,
            'webmd.com': 0.6, 'wikipedia.org': 0.4,
            'blogspot.com': 0.2, 'youtube.com': 0.2
        }
        self.academic_sites = {
            'researchgate.net', 'academia.edu', 'scholar.google.com',
            'arxiv.org', 'pubmed.ncbi.nlm.nih.gov', 'jstor.org'
        }
        self.sid = SentimentIntensityAnalyzer()
```

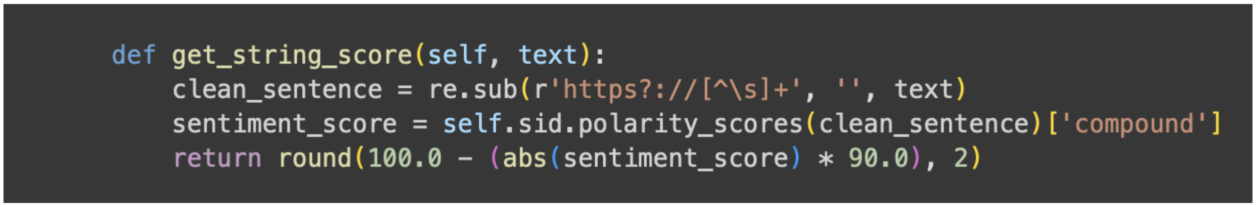
For each text, URLs are extracted and evaluated; the maximum domain credibility score is selected as the `link_score`, scaled to a 0–100 range. Research consistently highlights source reputation as a dominant factor in perceived trustworthiness. Castillo et al. (2011) showed that domain-level features (authority, reputation) are reliable predictors of information credibility on social platforms. Fogg et al. (2003) also demonstrated that users often infer trust based on institutional and domain signals rather than content specifics. By quantifying source trust, the algorithm aligns with established methods in computational credibility modeling.

Next, the text that is associated with the domain undergoes ‘sentiment evaluation’ using the VADER sentiment analyzer, yielding a `compound` score in the range [-1, 1]. The model translates this into a string score using:

string_score=100−(|sentiment|×90)string_score=100−(|sentiment|×90)

This transformation rewards neutral or objective tone (sentiment near zero) and penalizes extreme polarity (strongly positive or negative emotional tone). Horne & Adali (2017) found that false or low-credibility news tends to exhibit emotional, subjective, and hyperbolic language, while credible news is more linguistically neutral. Mitra & Gilbert (2015) showed that sentiment polarity is a strong signal in misinformation detection. Therefore, linguistic neutrality serves as an empirical proxy for informational reliability.

The `abs()` function ensures that both strong positive and strong negative sentiments reduce the score because extreme tone in either direction is less credible than neutrality.



```
def get_string_score(self, text):
    clean_sentence = re.sub(r'https?://[^\s]+', '', text)
    sentiment_score = self.sid.polarity_scores(clean_sentence)['compound']
    return round(100.0 - (abs(sentiment_score) * 90.0), 2)
```

After Hybrid Feature Engineering, we move on to Normalization and Model Training. Both `link_score` and `string_score` features are normalized using **Min–Max scaling**:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

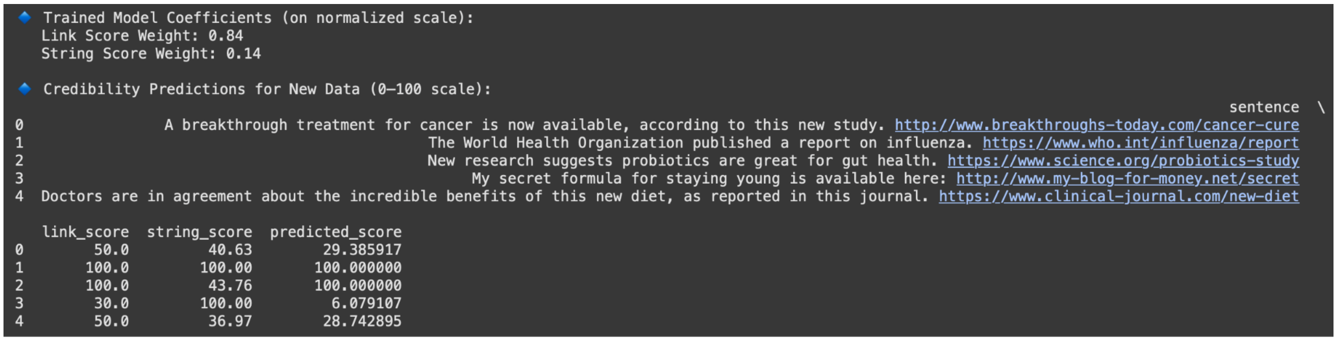
This ensures that both features contribute equally during model training and prevents dominance by any one feature. The target variable (`human_score`) is similarly normalized to improve regression stability.

A linear regression model is then trained on this normalized data, learning weights that best fit the human-assigned credibility ratings:

$$\hat{y} = w_1 \times \text{link\_score} + w_2 \times \text{string\_score} + b$$

Linear models provide transparent weight coefficients, enabling analysts to directly interpret which factor (source trust vs. tone neutrality) most influences credibility predictions.

The trained model then predicts credibility scores for unseen text data. In the trained model, `link_score` was given a weight of 0.84, while `string_score` was given a weight of 0.14. Predictions are then rescaled to the 0–100 range for interpretability and presentation. This makes results directly comparable to human assessments or institutional trust scores. For example, the string “A breakthrough treatment for cancer … www.breakthroughs-today.com” received a `link_score` of 50.0, a `string_score` of 40.63, and a `predicted_score` of 28.39.



```
◆ Trained Model Coefficients (on normalized scale):
  Link Score Weight: 0.84
  String Score Weight: 0.14

◆ Credibility Predictions for New Data (0–100 scale):
                                                                                                                                          sentence  \
0                A breakthrough treatment for cancer is now available, according to this new study. http://www.breakthroughs-today.com/cancer-cure
1                                              The World Health Organization published a report on influenza. https://www.who.int/influenza/report
2                                              New research suggests probiotics are great for gut health. https://www.science.org/probiotics-study
3                                                   My secret formula for staying young is available here: http://www.my-blog-for-money.net/secret
4  Doctors are in agreement about the incredible benefits of this new diet, as reported in this journal. https://www.clinical-journal.com/new-diet

   link_score  string_score  predicted_score
0        50.0         40.63        29.385917
1       100.0        100.00       100.000000
2       100.0         43.76       100.000000
3        30.0        100.00         6.079107
4        50.0         36.97        28.742895
```

The credibility scoring algorithm demonstrates several strengths that make it both practical and scientifically grounded. Its most significant advantage lies in its interpretability—each feature’s contribution to the overall credibility score can be traced directly through the model’s coefficients. By combining sentiment neutrality and domain trustworthiness, the model is consistent with well-established research in computational credibility and NLP-based misinformation detection. Its simplicity allows it to function effectively on small datasets, while the modular design means that additional credibility indicators, such as readability, bias lexicons, or metadata, can be seamlessly integrated in future iterations. Overall, the model balances theoretical rigor with computational efficiency, making it a strong foundation for explainable credibility assessment.

However, certain limitations should be acknowledged. The model’s reliance on a static domain dictionary restricts its adaptability to newly emerging or non-traditional information sources. Linguistically, sentiment analysis alone cannot capture subtleties such as sarcasm, figurative speech, or context-dependent credibility. The assumption of linear relationships between features and credibility may oversimplify complex interactions, while the lack of temporal or network-based analysis means that evolving credibility dynamics—such as misinformation spread over time—remain unaddressed. These constraints suggest that while the model is an effective proof-of-concept, additional enhancements are necessary for large-scale, real-world deployment.

In conclusion, this hybrid algorithm effectively merges domain trustworthiness and linguistic objectivity into a cohesive, interpretable credibility scoring framework. Its scientific grounding in computational trust modeling, sentiment analysis, and statistical regression supports both its reliability and transparency. While limited by static heuristics, it offers a strong foundation for scalable, data-driven credibility evaluation; bridging human judgment and algorithmic precision.