Model for Detecting Labels: True, Half-True, Mostly True, Mostly False, and False and Development of a Half-Truth Dataset

Akanksha Dadhich
M-Tech(RA) 23M0830
Computer Science and Engineering
Indian Institute of Technology
akankshadadhich@cse.iitb.ac.in

Abstract

In today's digital landscape, misinformation and deceptive content present growing challenges in distinguishing truth from falsehood. Among these, half-truths are particularly deceptive, as they blend factual elements with omissions, exaggerations, or misleading contexts to misrepresent the truth. This work introduces a novel framework for automatically detecting and categorizing deceptive statements into five distinct labels: True, Mostly True, Half-True, Mostly False, and False. The report first explores various techniques for converting true statements into half-truths and other categories, including omission of key information, ambiguity, context manipulation, selective presentation, paraphrasing, and adversarial attacks. These transformations enable the systematic generation of deceptive variations of true statements, which serve as the foundation for our dataset.

Additionally, the report details the creation of a new dataset by applying these techniques to true statements sourced from news articles. To evaluate the generated deceptive statements, we utilize GPT-4 for scoring various features, including Factual Accuracy, Deceptiveness, Coherence, Specificity, Emotional Tone, and Bias. We train an ensemble of models, including Support Vector Machines (SVM) and Gradient Boosting, on annotated data from the PolitiFact dataset to predict the final label based on these features. Our approach achieves high accuracy in labeling, effectively distinguishing between nuanced categories such as Mostly True and Half-True. This predictive model is then used to generate a novel dataset, enriching resources for future research in misinformation detection.

The proposed framework demonstrates its effectiveness in identifying and categorizing half-truths, contributing to ongoing efforts in combating misinformation. Our research highlights the importance of understanding half-truths as a subtle yet pervasive form of deception, emphasizing the need for robust methods to address this challenge in online content analysis and fact-checking.

1 Introduction

The rapid spread of misinformation, particularly in the form of half-truths, has become a significant challenge in today's digital landscape. A **half-truth** is a deceptive statement that includes some element of truth but intentionally omits or distorts critical details to mislead. These statements exploit cognitive biases and ambiguities in how information is presented, often influencing public opinion, shaping behavior, and undermining trust. Unlike outright falsehoods, half-truths operate subtly, requiring careful analysis to uncover their deceptive nature. This challenge is particularly prominent in digital media, where information spreads rapidly and distinguishing between true and deceptive statements becomes increasingly difficult. Misinformation campaigns often use half-truths to manipulate public perception, adding to the complexity of identifying misleading content.

The aim of this research is to develop a comprehensive model for detecting and categorizing statements as **True**, **Half-True**, **Mostly True**, **Mostly False**, or **False**. This categorization is crucial for combating misinformation in news, media, and online content. To achieve this, we explore a variety of techniques used to convert true statements into half-truths, such as omission of key information, selective presentation, paraphrasing, exaggeration, and ambiguity. These methods can subtly alter a statement's meaning while maintaining some degree of factual accuracy. For example, a statement such as:

- **Full Truth:** “The advertised product has been returned by a large number of customers, but it has also received some positive reviews.”

- **Half-Truth (Omission):** “The advertised product has received some positive reviews.”

The second version omits critical context about returns and negative feedback, misleadingly emphasizing the product’s positive reception.

To enhance the generation of half-truths, we employ advanced techniques such as adversarial attacks, which subtly alter key phrases or sentiments to misrepresent the truth. Techniques like generalization, omission, and selective presentation enable the creation of statements that distort or oversimplify facts, resulting in misleading narratives.

To address this problem, we leverage publicly available datasets, such as the **Real and Fake News** dataset and the **PolitiFact** dataset. The **Real and Fake News** dataset consists of news articles labeled as real or fake, while the **PolitiFact** dataset contains fact-checked statements, offering valuable insights into the accuracy and credibility of various claims. These datasets are essential for training machine learning models that can classify statements based on their factual accuracy, helping to identify both true and deceptive content.

The primary motivation for generating **deceptive statements** is to better understand how misinformation, especially in the form of half-truths, is crafted and propagated. By examining how seemingly truthful statements can be manipulated through techniques such as omission, selective presentation, and exaggeration, we aim to develop tools to detect and categorize these subtle forms of deception. This is critical for combating misinformation, particularly in news, media, and online content.

Contributions

This research contributes to the detection and analysis of half-truths by:

- Developing a comprehensive model for classifying statements into five categories: True, Half-True, Mostly True, Mostly False, and False.

- Creating a novel dataset of deceptive statements derived from real news articles, annotated to highlight the elements of truth and deception.

- Exploring machine learning techniques such as Support Vector Machines (SVM) and Gradient Boosting to identify the most effective methods for detecting deceptive statements.

This research involves creating a new dataset of deceptive statements derived from real news articles. These statements are passed through a machine learning model trained on a curated dataset from **PolitiFact**, which evaluates statements based on criteria such as factual accuracy, deceptiveness, coherence, specificity, and emotional tone. This framework assigns scores to statements and categorizes them into one of the five labels. The dataset helps to ensure that the statements reflect common forms of deception encountered in everyday media consumption, and it serves as a valuable resource for further research in misinformation detection and classification.

The machine learning framework explores various models, including Support Vector Machines (SVM) and Gradient Boosting, to determine which provides the best accuracy in predicting the labels. These models are evaluated based on their ability to utilize features such as factual accuracy, emotional tone, bias, and coherence to effectively categorize statements as True, Half-True, Mostly True, Mostly False, or False. By comparing the performance of these techniques, the research aims to identify the most suitable model for accurately capturing the subtle nuances of half-truths and other deceptive statements.

Ultimately, this research aims to provide a robust tool for detecting and analyzing the subtleties of misinformation, particularly in the form of half-truths. By deepening our understanding of how these deceptive statements are constructed and recognized, we contribute to broader efforts to enhance trust and credibility in information sources, especially in the digital era. The findings of this research underscore the need for more sophisticated tools to detect deceptive content, where misinformation can easily spread and influence public opinion.

2 Background and Related Work

PolitiFact has emerged as a leading platform in the fact-checking domain, renowned for its Truth-O-Meter system, which evaluates the accuracy of statements made by political figures, public personalities, and pundits. Established in 2007 by the Tampa Bay Times, PolitiFact employs a dedicated team of journalists and researchers who thoroughly investigate claims. Their rigorous methodology involves examining evidence from diverse sources and employing expert analysis to verify statements. PolitiFact's approach prioritizes transparency, providing detailed explanations and citations for each verdict, enabling readers to assess the evidence independently. Ratings range from "True" to "Pants on Fire," indicating the degree of accuracy.

Advances in natural language processing (NLP) and machine learning have transformed the field of misinformation detection. The introduction of the Bidirectional Encoder Representations from Transformers (BERT) model by Google in 2018 has been a significant milestone. BERT employs a transformer-based architecture to process text bidirectionally, capturing intricate linguistic and semantic nuances. Its performance in tasks like sentiment analysis, named entity recognition, and question answering has been state-of-the-art. Leveraging BERT's contextual language understanding, researchers have fine-tuned it on specialized datasets to enhance its capability for detecting deceptive content.

Datasets such as SentimentalLiar (Upadhayay and Behzadan, 2020) have introduced a comprehensive framework for sentiment analysis, categorizing text into five emotion classes: joy, sadness, anger, fear, and surprise. The dataset also includes sentiment scores, numerical indicators of the positive or negative tone of a text. These features allow researchers to assess the emotional undertones and categorize text as positive, negative, or neutral. Such tools are valuable for analyzing how sentiment influences the spread and perception of misinformation.

Convolutional Neural Networks (CNNs) have also proven effective for tasks like text classification, including misinformation detection. By applying convolutional filters, CNNs excel at identifying local patterns and features in textual data, learning hierarchical representations that help detect subtle indicators of deception. Similarly, Recurrent Neural Networks (RNNs) have shown significant promise in analyzing sequential data, such as text or time-series information. Unlike traditional neural networks, RNNs model temporal dependencies, making them suitable for tasks involving sequential inputs.

A notable example of RNNs' effectiveness is their application in detecting rumors on social media platforms, as demonstrated by Ma et al. (2016). Their research utilized RNNs to model the temporal dynamics of rumor propagation in microblogs, identifying patterns indicative of misinformation. This approach highlights the potential of RNNs in understanding how deceptive information spreads, offering critical tools to combat misinformation in online networks.

Despite significant progress, automatic fact-checking tools face limitations that restrict their broader adoption by professional fact-checking organizations like PolitiFact. Many studies rely on crowdsourced claims, which often fail to reflect the complexities encountered in real-world fact-checking. Other research depends on curated "gold-standard" evidence or unrestricted retrieval of related content, which risks retrieving articles that merely discuss claims rather than providing independent evidence. Systems that retrieve evidence from diverse, unstructured sources remain underdeveloped, as noted in previous work (Ferreira and Vlachos (2016); Alhindi et al. (2018); (Atanasova et al., 2020)).

These challenges underscore the need for more robust systems that address the intricacies of real-world misinformation and offer scalable solutions for fact-checking in diverse and dynamic environments.

3 Methodology

Evaluation Model (M Model) Development

The M Model was designed as an evaluative model capable of analyzing and scoring statements based on various criteria to assess truthfulness, coherence, and the degree of potential deception. This model was developed using the PolitiFact dataset, which contains labeled claims verified by professional fact-checkers, making it a reliable source of ground truth for the training phase.

To train the M Model, a supervised learning approach was employed, utilizing only the PolitiFact dataset. Labels from PolitiFact, such as *"True," "Mostly True," "Half True," "Mostly False,"* and *"False,"* provided the ground truth for model optimization. The training process aimed to build a classifier that could reliably determine the degree of truthfulness and deception within statements by aligning with PolitiFact's annotations.

Evaluation Prompt and Scoring Criteria

To evaluate each statement, GPT-4 was used to score the statements based on nine specific criteria, each on a 0 to 1 scale. This approach provided structured feedback on each statement's alignment with factual accuracy and other qualitative metrics, such as specificity and coherence. The final label was determined by GPT-4, which generated a label based on aggregated criteria scores. Statements, where the GPT-4 label matched the PolitiFact label, were retained for further analysis, ensuring that only high-confidence data points contributed to model learning.

Scoring Criteria

- **Factual Accuracy (0 to 1):**
  - **Definition:** Measures how accurately a statement reflects evidence, identifying factual alignments or inaccuracies.
  - **Example:** A statement like "Vaccines completely eliminate COVID-19 risk" would score low (e.g., 0.2) due to its inaccuracy.
  - **Importance:** High factual accuracy is crucial for preventing misinformation and misinformed actions.

- **Deceptiveness (0 to 1):**
  - **Definition:** Evaluates how misleading a statement might be through exaggeration, omission, or implication.
  - **Example:** A minor funding cut described as having drastic effects might receive a high deceptiveness score (e.g., 0.8).
  - **Importance:** Captures subtle manipulations that, while not outright false, could mislead the audience.

- **Coherence (0 to 1):**
  - **Definition:** Assesses logical flow and clarity in the statement.
  - **Example:** A convoluted statement on healthcare policy could score low on coherence (e.g., 0.3) if it lacks logical structure.
  - **Importance:** Coherent statements ensure that information is comprehensible and reliable.

- **Specificity (0 to 1):**
  - **Definition:** Rates the level of detail, distinguishing specific details from vagueness.
  - **Example:** "Many benefit from healthcare reform" would score low (e.g., 0.2) due to lack of specific detail.
  - **Importance:** High specificity reduces ambiguity, providing clearer insights.

- **Emotional Tone (0 to 1):**

  - **Definition:** Measures the degree to which a statement appeals to emotion.
  - **Example:** Statements with charged language like “cruel policy” might score high (e.g., 0.9), aiming to provoke strong emotions.
  - **Importance:** Emotional language can skew objectivity and influence readers’ perception of facts.

- **Bias (0 to 1):**
  - **Definition:** Assesses if a statement favors a particular viewpoint.
  - **Example:** A claim like “government intervention is the only solution to environmental issues” reflects bias and may score high (e.g., 0.8).
  - **Importance:** Ensuring low bias allows for balanced, multi-perspective information.

- **Scope/Generality (0 to 1):**
  - **Definition:** Evaluates whether a statement is overly broad or narrowly focused.
  - **Example:** “Politicians always act in their own interest” would score high (e.g., 0.9) for generalization.
  - **Importance:** Reducing overgeneralizations results in statements that better reflect nuanced realities.

- **Temporal Consistency (0 to 1):**
  - **Definition:** Assesses time relevance to ensure accuracy within the context.
  - **Example:** A claim like “Unemployment is at an all-time low” may score low (e.g., 0.3) if it’s outdated.
  - **Importance:** Ensures that statements accurately reflect current conditions, preventing misinterpretation over time.

- **Out of Context or Ambiguity (0 to 1):**
  - **Definition:** Identifies statements lacking context or containing ambiguous language.
  - **Example:** “The economy grew by 5%” could score low (e.g., 0.4) without specifying a timeframe or sector.
  - **Importance:** Providing context enhances transparency and reduces the risk of misunderstanding.

Model Training and Optimization

With the scoring criteria established, the M Model was trained using a range of algorithms, including SVM, Gradient Boosting, and Neural Networks. Each algorithm was evaluated for its performance in accurately classifying statements based on truthfulness, coherence, and other criteria. The objective was to achieve a model that could reliably predict the truthfulness label with a high level of accuracy. Following training, the model achieved an accuracy of 85%.

The key steps in model training and optimization included:

- **Feature Selection:** The model incorporated scores from each of the nine criteria, enabling it to recognize patterns in truthful and deceptive statements.

- **Cross-Validation:** A validation process ensured that the model generalized well to unseen data, avoiding overfitting.

- **Parameter Tuning:** Parameters were adjusted across algorithms to find the best-performing configuration, balancing accuracy with computational efficiency.

- **Final Model Choice:** The model achieving 85% accuracy was selected as the final candidate, providing an optimal balance of accuracy and interpretability.

True Label and Score Generation from PolitiFact Dataset

To ground the M Model in truth-evaluation standards, the true labels for each statement were sourced from PolitiFact. These labels served as a gold standard, anchoring the model's predictions to a verified truth and falsehood benchmark. True labels from the PolitiFact dataset were used as ground truth to supervise the training process. Each statement from PolitiFact was assigned a label (True, Mostly True, Half True, Mostly False, or False). ChatGPT was used to score each statement based on the nine criteria, providing deeper insights into the degree of deceptiveness, coherence, and specificity for each statement. Only statements where GPT-4's generated label aligned with PolitiFact's true label were retained, reinforcing the model's reliability.

Data Filtering for Confidence in Scores

A rigorous data filtering process was implemented to improve the M Model's reliability. Statements where GPT-4's label predictions aligned with PolitiFact's labels were considered high-confidence data points and retained for further training. This filtering step ensured that the model focused on high-quality examples, reinforcing the dataset's integrity and improving the final model's consistency and trustworthiness in evaluating truthfulness. The data filtering process aimed to enhance confidence in the scores generated by GPT-4.This filtering process reduced noise in the dataset and ensured that only high-quality, confidently labeled examples were used to train the M Model.

Half-Truth Dataset Creation

Half-Truths as Deceptive Statements

A half-truth is inherently deceptive because it manipulates the listener's perception by combining elements of truth with omissions, ambiguities, or distortions. The deceptive nature arises from the speaker's intent to mislead, evade blame, or misrepresent reality while maintaining some level of plausibility.

Why Half-Truths are Deceptive:

- **Difficult to Identify:** Determining whether a statement is a half-truth often requires significant contextual knowledge, making it challenging even for informed listeners.
  - **True Claim:** "The Jammu and Kashmir Reorganisation Act, 2019 was passed by the parliament, enacting the division of the state into two union territories."
  - **Half-Truth:** "Jammu and Kashmir was reorganized into new territories by an act of parliament."
  - **Explanation:** This version omits the year and act's name, which changes the clarity of the event and manipulates the listener's perception (Context Manipulation).

- **Reliance on Trust:** Listeners unfamiliar with the topic or trusting the speaker are more likely to believe half-truths.
  - **Statement:** "India won chess Olympiad because of Gukesh Dommaraju."
  - **Explanation:** While Gukesh played a significant role, this oversimplifies the contributions of other team members, misleading listeners (Selective Omission).

- **Easier to Detect with Familiarity:** When the listener has expertise in the subject, identifying omitted details becomes easier, revealing the intent to deceive.
  - **True Claim:** "In 2020, OpenAI released GPT-3, a language model with 175 billion parameters, designed to generate human-like text based on input prompts."
  - **Half-Truth:** "OpenAI released GPT-3, an advanced AI model."
  - **Explanation:** Omits critical technical details and the release year, reducing the listener's ability to fully understand the statement (Ambiguity).

Techniques for Generating Deceptive Statements

1. Omission: This involves leaving out critical information to make a statement appear more favorable.

- **Example:** "Our new product has received rave reviews!"
  Deception: Omits that only a small percentage of reviews were positive, while most were negative.
  Truth: A few reviews praised the product, but the majority criticized it.

- **Example:** "Our app has been downloaded over a million times!"
  Deception: Ignores that most downloads occurred years ago, and the app is no longer widely used.
  Truth: Active user base is significantly smaller.

- **Example:** "Our candidate has always fought for workers' rights."
  Deception: Omits that the candidate voted against key workers' rights legislation in the past.
  Truth: The candidate supports workers' rights now, but not consistently.

- **Example:** "We've partnered with industry leaders to improve our services."
  Deception: Omits that the partnerships are limited to minor collaborations.
  Truth: Partnerships have minimal impact on core services.

- **Example:** "This medication has been shown to improve health outcomes."
  Deception: Omits that improvements were minor and only in a small subset of patients.
  Truth: Benefits are limited to specific cases.

2. Exaggeration: Statements that amplify truth to an unreasonable or unsubstantiated degree.

- **Example:** "This is the most revolutionary product in the entire tech industry!"
  Deception: Exaggerates the significance of the product, implying unmatched innovation.
  Truth: The product has some innovative features, but it's not industry-leading.

- **Example:** "Our solution eliminates 100% of security threats."
  Deception: Overstates the product's effectiveness.
  Truth: It reduces certain threats but doesn't eliminate all.

- **Example:** "Our CEO is a global icon who has transformed multiple industries."
  Deception: Exaggerates achievements, implying a far-reaching impact.
  Truth: The CEO has had notable success in one industry.

- **Example:** "This diet plan guarantees weight loss for everyone!"
  Deception: Overstates the plan's effectiveness.
  Truth: Weight loss depends on individual adherence and metabolism.

- **Example:** "This car is the fastest on the market."
  Deception: Exaggerates performance, as faster cars exist.
  Truth: It's among the faster models in its price range.

3. Understatement: Downplaying significant aspects to make them seem less impactful.

- **Example:** "We've made a few small updates to our product."
  Deception: Downplays a complete overhaul.
  Truth: The product was redesigned with major feature upgrades.

- **Example:** "The company had a slight drop in revenue last quarter."
  Deception: Minimizes a significant 30% revenue loss.
  Truth: The drop represents substantial financial challenges.

- **Example:** "There's been some minor criticism of our policies."
  Deception: Downplays widespread public backlash.
  Truth: Policies have faced significant opposition.

- **Example:** "We've adjusted prices slightly to account for inflation."
  Deception: Downplays a 20% price increase.
  Truth: The price hike is substantial and beyond inflation.

- **Example:** "Our project experienced a brief delay."
  Deception: Minimizes a 6-month delay.
  Truth: The delay significantly affected timelines.

4. Alteration of Facts: Modifying factual elements to misrepresent reality.

- **Example:** "Our product was named the best in the industry by TechWorld last month!"
  Deception: Alters the fact; it was only listed as a notable product.
  Truth: The product was featured but not awarded.

- **Example:** "Our charity helped 1,000 families last year."
  Deception: Changes the context; it provided one-time support, not ongoing aid.
  Truth: Assistance was limited and not sustained.

- **Example:** "The team won the championship in 2023!"
  Deception: Omits that it was a regional, not national, championship.
  Truth: The achievement was significant but on a smaller scale.

- **Example:** "The study proves our product cures migraines."
  Deception: Alters findings; the study shows symptom reduction, not a cure.
  Truth: The product alleviates some symptoms.

- **Example:** "Our event broke attendance records!"
  Deception: Alters context; it broke a record for a specific niche, not overall attendance.
  Truth: Record-breaking within a smaller category.

5. Over-Representation of Numbers: Using inflated figures or misleading numerical claims.

- **Example:** "Thousands of users have switched to our product!"
  Deception: Inflates actual numbers; only a few hundred users switched.
  Truth: Uptake is growing but modest.

- **Example:** "We've cut costs by 50% this year!"
  Deception: Applies only to a small department, not overall costs.
  Truth: Savings were limited to specific areas.

- **Example:** "Over 90% of our customers recommend us!"
  Deception: Based on a small, cherry-picked survey sample.
  Truth: Broader surveys show lower satisfaction.

- **Example:** "Our product lasts five times longer than competitors'!"
  Deception: Compares only to a low-quality competitor.
  Truth: Longevity is average among premium products.

- **Example:** "We've tripled our workforce in the last year!"
  Deception: Original workforce was very small, making the increase less impressive.
  Truth: Growth is real but modest.

6. Generalization: Making vague, sweeping claims without specific evidence to back them up.

- **Example:** "Our product has improved user satisfaction across all areas."
Deception: Generalizes improvements without specifying which areas.
Truth: Satisfaction improved slightly in usability but not in durability.

- **Example:** "Everyone is switching to our service."
Deception: Implies a universal trend, while only a niche audience is switching.
Truth: A specific demographic is switching.

- **Example:** "Our policy has been a success for all stakeholders."
Deception: Overgeneralizes without mentioning groups that faced negative impacts.
Truth: Success was limited to some stakeholders.

- **Example:** "Our program works for any business."
Deception: Overstates applicability; it works only for small to medium businesses.
Truth: Larger businesses require additional customization.

- **Example:** "This diet plan is perfect for everyone."
Deception: Overgeneralizes; it may not suit people with certain health conditions.
Truth: It is effective for specific groups.

7. Context Manipulation: Presenting a statement without the full context to create a false impression.

- **Example:** "Our product is 50% faster than the previous version."
Deception: Applies only to certain tasks, not overall performance.
Truth: Performance improvements are limited to specific functions.

- **Example:** "Our team's sales grew by 40% this year."
Deception: Ignores that the growth was from a low baseline.
Truth: Sales were initially minimal, making the growth less impressive.

- **Example:** "This medication has no reported side effects."
Deception: Ignores that the medication is new, with insufficient testing.
Truth: Long-term side effects are unknown.

- **Example:** "Our company has the highest market share in the industry."
Deception: The company leads in a small sub-category, not the overall market.
Truth: Market share dominance is niche-specific.

- **Example:** "Our competitor's product was recalled."
Deception: Leaves out that the recall was voluntary and limited in scope.
Truth: Recall was minor and resolved quickly.

8. Ambiguity: Using vague language that can be interpreted in multiple ways.

- **Example:** "Our product provides unmatched performance."
Deception: Does not specify what "unmatched" refers to.
Truth: Performance is strong in specific scenarios, but not overall.

- **Example:** "We guarantee results!"
Deception: Ambiguous about the nature of the results or conditions for the guarantee.
Truth: Guarantee may come with significant conditions.

- **Example:** "Our services are used by leading companies."
Deception: Vague about what "leading" means or the level of usage.
Truth: Used by a few leading companies, but not extensively.

- **Example:** "This supplement supports a healthy lifestyle."

Deception: Ambiguous about what "supports" entails.
Truth: May contribute indirectly to health under certain conditions.

- **Example:** "We are the industry's go-to solution."
  Deception: Ambiguous about which industry or scope.
  Truth: Popular only in specific regions or segments.

9. Quantifier Shift: Changing quantifiers to make a statement seem more impactful.

- **Example:** "Many users prefer our product."
  Deception: "Many" is less definitive than "most," making the claim misleading.
  Truth: A portion of users prefer the product, but it's not the majority.
- **Example:** "Some studies suggest our product is effective."
  Deception: Implies broad scientific support when only a few studies exist.
  Truth: Evidence is limited and not conclusive.
- **Example:** "A large percentage of users report satisfaction."
  Deception: Leaves out that the percentage is relative to a small sample.
  Truth: Satisfaction is high among a niche audience.
- **Example:** "Dozens of experts endorse our approach."
  Deception: "Dozens" is less impactful than "hundreds" or "most."
  Truth: Endorsements exist but are not overwhelming.
- **Example:** "Our app has millions of downloads."
  Deception: Suggests ongoing popularity; downloads may have peaked years ago.
  Truth: Active user base is much smaller.

10. Selective Comparison: Highlighting favorable comparisons while ignoring relevant context.

- **Example:** "Our product is 20% cheaper than the leading competitor."
  Deception: Ignores differences in features or quality.
  Truth: Lower price comes with fewer features.
- **Example:** "Our car is more fuel-efficient than Brand X."
  Deception: Compares against the least fuel-efficient model.
  Truth: Efficiency is average compared to top competitors.
- **Example:** "Our service has faster response times than others."
  Deception: Applies only to specific hours or locations.
  Truth: Response times vary widely.
- **Example:** "Our phone has a larger screen than the iPhone."
  Deception: Ignores differences in resolution and quality.
  Truth: Screen size is larger, but quality may not match.
- **Example:** "We offer more features than any competitor."
  Deception: Exaggerates; features may be superficial.
  Truth: Feature count is high, but not all are useful.

11. False Equivalence: Drawing inappropriate parallels to make a statement seem credible.

- **Example:** "Choosing our product is like choosing clean energy—it's the ethical choice."
  Deception: Equates the product with a universally valued cause.
  Truth: Ethical implications are subjective.

- **Example:** "Our product is as essential as your smartphone."
  Deception: Equates a niche product with a universal necessity.
  Truth: The product serves a specialized need.

- **Example:** "Our program is as effective as hiring a personal coach."
  Deception: Overstates the effectiveness of the program.
  Truth: It provides guidance, but not personalized coaching.

- **Example:** "Using our service is like saving the planet!"
  Deception: Exaggerates environmental impact.
  Truth: The service has limited eco-benefits.

- **Example:** "Our app is like having a personal assistant."
  Deception: Suggests human-like capabilities.
  Truth: It automates simple tasks.

12. Misleading Cause and Effect: Implying a direct relationship between two unrelated events.

- **Example:** "Thanks to our product's launch, customer satisfaction has skyrocketed!"
  Deception: Satisfaction may have improved due to unrelated factors.
  Truth: The product's launch coincided with other improvements.

- **Example:** "Sales improved because of our marketing campaign."
  Deception: Sales might have improved due to seasonal demand.
  Truth: Campaign impact is unclear.

- **Example:** "Our software reduced cyber-attacks by 50**Deception:** Reduction could be due to unrelated external factors.
  Truth: Software may have contributed partially.

- **Example:** "Our investment program led to record profits."
  Deception: Record profits may result from broader economic trends.
  Truth: Program played a minor role.

- **Example:** "After using our service, customers report better health."
  Deception: Health improvements may be unrelated.
  Truth: Service may have minor benefits.

13. Emotional Appeal: Using emotions to obscure factual evaluation.

- **Example:** "If you care about your family's safety, you need our product!"
  Deception: Plays on fear to encourage purchase.
  Truth: Product may enhance safety, but it's not essential.

- **Example:** "Don't let your loved ones suffer—try our solution today."
  Deception: Exploits guilt to influence behavior.
  Truth: Solution may not suit everyone.

- **Example:** "Show your love with our premium gift set!"
  Deception: Suggests that love requires purchasing the product.
  Truth: The gift set is optional.

- **Example:** "Choose our service and join the fight for justice."
  Deception: Aligns the service with a noble cause.
  Truth: Service has minimal relation to justice.

- **Example:** "Imagine the heartbreak of missing out—buy now!"
  Deception: Exaggerates consequences of inaction.
  Truth: The product is not time-sensitive.

Score Generation of the Statements Produced

Statements derived from real news data were modified using techniques like ambiguity, bias, and selective presentation, then evaluated on dimensions such as Factual Accuracy, Deceptiveness, Coherence, Specificity, Emotional Tone, and Context. Scores for each dimension were assigned based on the alignment with evidence and clarity of the claims using GPT-4 mini.

Generating Final Labels Using the M Model

The trained M Model used the generated scores to predict final labels—True, Mostly True, Half True, Mostly False, or False. Labels were determined by thresholds, where high factual accuracy and low deceptiveness predicted “True," while low accuracy and high deceptiveness predicted “False." This automated process ensured consistent, scalable evaluation of statement credibility.

4 Experiments and Results

4.1 Training the M Model

4.1.1 Experiment 1: Initial Model Training with PolitiFact Data

Setup:

- **Dataset:** PolitiFact (250 instances, 50 instances per class: *True*, *Half-True*, *Mostly-True*, *Mostly-False*, *False*).
- **Features:** Deceptiveness, Factual Accuracy, Coherence (scores generated via GPT-4 prompts).
- **Models Evaluated:** Linear Regression, Decision Tree, SVM, Random Forest, Gradient Boosting Classifier.
- **Target:** Prediction of gold labels (*True*, *Half-True*, *Mostly-True*, *Mostly-False*, *False*).

Results:

- **Linear Regression:** Failed to fit the data effectively. Example prediction for Factual Accuracy = 0.2, Deceptiveness = 0.7, Coherence = 1 gave a score inconsistent with the ground truth.
- **Decision Tree:** Accuracy: **58%**. Precision, recall, and F1-scores were inconsistent across classes, with Class 0 (*True*) performing poorly.
- **SVM:** Accuracy: **58%**, with minor improvements in recall and F1-scores over the Decision Tree.
- **Random Forest:** Accuracy: **64%**, with better overall balance between precision and recall.
- **Gradient Boosting Classifier:** Accuracy: **68%**, the highest among tested models, with consistent performance across all classes.

Analysis:

- Gradient Boosting performed best due to its ability to handle imbalanced data and nonlinear interactions among features.
- Feature importance analysis suggested that *Coherence* contributed less than expected, potentially introducing noise.

4.1.2 Experiment 2: Feature Engineering to Improve Model Accuracy

Setup:

- **Dropped:** The *Coherence* feature based on low contribution from Experiment 1.
- **Added five new features:**
    1. **Specificity:** Measures detail level of the claim.
    2. **Emotional Tone:** Assesses sentiment in the language.
    3. **Scope/Generality:** Evaluates the breadth of the statement.
    4. **Temporal Consistency:** Checks alignment with the described timeframe.
    5. **Out-of-Context or Ambiguity:** Rates clarity and contextual alignment.

Results:

- **Decision Tree:** Accuracy: **70%**. Improved precision and recall, particularly for Classes 3 (*Mostly-False*) and 4 (*False*).
- **Random Forest:** Accuracy: **70%**, consistent across all classes.
- **Gradient Boosting Classifier:** Accuracy: **70%**, retained top performance but with reduced variance in class-wise metrics.
- **SVM:** Accuracy: **74%**, highest improvement among models. Macro F1-score increased due to better handling of Classes 1 (*Half-True*) and 3 (*Mostly-False*).

Analysis:

- Removing *Coherence* eliminated noise, while new features enriched the model's ability to discern nuanced patterns, especially in subjective aspects like emotional tone and ambiguity.
- SVM outperformed others by leveraging these high-dimensional features effectively.

4.1.3 Experiment 3: Data Augmentation with GPT-4 Mini

Setup:

- **Generated:** 250 synthetic instances (50 per class) using GPT-4 Mini to supplement PolitiFact data.
- **Feature Set:** Same as Experiment 2.

Results:

- **SVM:** Accuracy: **72%**, consistent with real data performance but with slight degradation in precision for Class 1 (*Half-True*).
- **Random Forest:** Accuracy: **68%**, comparable to Experiment 1, indicating no significant benefit from augmented data.
- **Decision Tree:** Accuracy: **66%**, minor drop from Experiment 2, likely due to overfitting on the synthetic data.
- **Gradient Boosting Classifier:** Accuracy: **70%**, stable despite data augmentation.

Analysis:

- Augmented data did not significantly enhance accuracy, likely due to the limited diversity in GPT-4 Mini's generated instances.
- The stability of Gradient Boosting Classifier and SVM suggests robustness to noisy data.
- This highlights the importance of high-quality data over quantity in this task.

4.1.4 Experiment 4: GPT-4 Mini for Data Processing and SVM

Setup: Retained data points where GPT-4 Mini predictions matched ground truth.

Results:

- **SVM:** Accuracy: 0.85, with high recall for label 0 and f1-scores above 0.80 for most classes.

Insights: Filtering based on GPT-4 Mini's accurate predictions improved model accuracy and highlighted the importance of consistent pre-processing.

4.1.5 Experiment 5: Enlarging Dataset and Feature Optimization

Setup: Increased dataset size and identified optimal feature combinations for SVM.

Results:

- **SVM:** Accuracy: 0.896, with improved recall for labels 1 and 4, and consistently perfect precision for Class 4.

Insights: Larger dataset size improved generalization. Careful feature selection (e.g., Factual Accuracy, Deceptiveness, Emotional Tone, Bias) boosted accuracy.

4.1.6 Experiment 6: Incorporating Mistral and Evaluating Multiple Models

Setup: Explore model performance using Mistral-generated scores.

Results:

- **SVM:** Accuracy: 0.78, best recall for label 3 but imbalanced performance for labels 0 and 4.
- **Random Forest:** Accuracy: 0.68, varying f1-scores across classes.
- **Decision Tree:** Accuracy: 0.70, strong recall for 0, but inconsistent metrics.
- **Gradient Boosting:** Accuracy: 0.70, balanced but not significantly better than SVM.
- **LightGBM:** Accuracy: 0.68, varied precision and recall.
- **Neural Networks:** Training Accuracy: 0.83, Test Accuracy: 0.79.

Insights: SVM showed balanced accuracy and overall performance. Neural networks exhibited potential but some overfitting.

4.2 Fine-tuning with PolitiFact and Newly Created Dataset

Two experiments were conducted to evaluate the performance of fine-tuned language models (BERT and XLM-RoBERTa) on a multi-class classification task with five labels: *True, Mostly-True, Half-True, Mostly-False, False*. The experiments aimed to analyze the impact of training on a newly created dataset versus domain-specific data from PolitiFact.

4.2.1 Experimental Setup

Experiment 1: Models were fine-tuned using a dataset of 800 instances created from real news articles using techniques such as paraphrasing and omission. Evaluation was conducted on an 80-instance test set from the PolitiFact dataset.

Experiment 2: Models were fine-tuned using 800 instances from the PolitiFact dataset and tested on the same 80-instance test set from PolitiFact.

Performance metrics such as accuracy, precision, recall, and F1-score were used for evaluation, along with confusion matrices for detailed analysis.

4.2.2 Results and Analysis

Experiment 1: Training with Real-News Dataset

BERT:

- **Accuracy:** 38%
- **Observations:**
  - The model exhibited imbalanced performance across classes, with the highest recall for *Mostly-True* (0.75) and *Half-True* (0.62), indicating better recognition of moderately deceptive statements.
  - The *False* label showed high precision (0.67) but low recall (0.12), suggesting precision in predictions but a failure to capture all instances.

XLM-RoBERTa:

- **Accuracy:** 41%
- **Observations:**
  - Performance was more balanced than BERT, with improved precision and recall for *Mostly-True*, *Half-True*, and *Mostly-False*.
  - The *False* label had high precision (0.83) but lower recall (0.31), similar to BERT.

Insights: Fine-tuning on the real-news dataset produced models with moderate performance on PolitiFact data, indicating difficulty in adapting to PolitiFact-specific nuances.

Experiment 2: Training with PolitiFact Dataset

BERT:

- **Accuracy:** 44%
- **Observations:**
  - Stronger recall for *Mostly-False* (0.44) and *False* (0.62), reflecting improved recognition of highly deceptive statements.

- The *Mostly-True* label remained challenging, with poor precision (0.21) and recall (0.19).

XLM-RoBERTa:

- **Accuracy:** 44%
- **Observations:**
  - Strong performance for *True* and *False* labels, with F1-scores of 0.56 and 0.59, respectively.
  - Confusion between *Mostly-False* and *False* was reduced compared to Experiment 1.

Insights: Models trained on PolitiFact data performed better in recognizing domain-specific patterns, with comparable performance between BERT and XLM-RoBERTa.

4.2.3 Comparative Analysis

Performance Across Models:

Metric	BERT (Real Data)	XLM-R (Real Data)
Accuracy	38%	41%
Macro F1	0.33	0.41

Table 1: Model Performance on Real Data

Metric	BERT (PolitiFact)	XLM-R (PolitiFact)
Accuracy	44%	44%
Macro F1	0.43	0.43

Table 2: Model Performance on PolitiFact Data

Label-Wise Insights:

- The *False* label consistently achieved high precision, reflecting strong model confidence in predictions.
- *Mostly-True* and *Half-True* were the most challenging labels, with frequent confusion due to semantic overlaps.
- PolitiFact-trained models showed enhanced recognition of *Mostly-False* and *False*, highlighting the importance of domain-specific data.

Error Analysis:

- *Confusion:* Significant confusion between neighboring labels, such as *True* and *Mostly-True*, or *Mostly-False* and *False*, was observed across all experiments, reflecting the nuanced nature of these categories.
- *Data Dependency:* Models trained on PolitiFact data better captured its stylistic and semantic nuances, while the real-news dataset led to generalized learning that struggled with PolitiFact-specific features.

5 Conclusion

Based on the experiments, the SVM model trained in Experiment 5, which included an expanded dataset and optimized feature selection, was selected for deployment. This decision was guided by several key factors:

- **Highest Accuracy:** The SVM model in Experiment 5 achieved the highest accuracy, reaching 0.896, which reflects strong predictive capability across diverse statement classes.
- **Balanced Performance Across Labels:** This model demonstrated consistently balanced performance across all labels, ensuring that no single class was disproportionately under- or over-represented in predictions.
- **Effective Feature Selection:** The inclusion of optimized features such as Factual Accuracy, Deceptiveness, Emotional Tone, Bias, Scope/Generality, and Temporal Consistency contributed to the model's efficiency and reliability. This feature selection process minimized noise and emphasized factors most relevant to accurate classification.

In summary, the SVM model from Experiment 5 was chosen for its accuracy, balanced class performance, and optimized feature set. This model is expected to generalize well on unseen data, making it a reliable choice for

Class	Precision	Recall	F1-Score	Support
0	0.88	0.88	0.88	16
1	1.00	0.54	0.70	13
2	0.83	1.00	0.91	15
3	0.84	1.00	0.91	16
4	1.00	1.00	1.00	17
Accuracy	0.90 (Overall Accuracy)			
Macro Avg	0.91	0.88	0.88	77
Weighted Avg	0.91	0.90	0.89	77

Table 3: SVM Model Performance Metrics

deployment in applications that require accurate and nuanced classification of statement truthfulness.

Models trained on PolitiFact data performed better on PolitiFact test data, demonstrating the importance of domain-specific fine-tuning. XLM-RoBERTa consistently outperformed BERT when trained on a diverse dataset, likely due to its enhanced pretraining.

6 Future Work and Limitations

Moving forward, several avenues for future research and improvement can be explored:

6.1 Limitations

- **Dataset Source Mismatch:** The training data used for fine-tuning was created using statements generated from Indian news sources, while the testing data for fine-tuning was based on the U.S.-centered PolitiFact dataset. This mismatch may contribute to lower accuracy after fine-tuning, as the model may struggle with cultural or linguistic nuances specific to U.S. or Indian news contexts.

  Future work could focus on augmenting training data with more nuanced, domain-specific examples and employing techniques to reduce confusion between neighboring classes, such as hierarchical classification or fine-grained feature engineering.

6.2 Future Work

To address these limitations and further improve the model's capabilities, we propose several future directions:

- **Cross-Regional Training Dataset:** Expanding the dataset to include statements from both Indian and U.S. news sources, as well as additional regions, could enhance the model's cross-cultural adaptability. A more diverse dataset would help the model better generalize across different contexts and news styles.

- **Fine-Tuning Larger Language Models (LLMs):** Fine-tuning advanced language models, such as GPT-4 or LLaMA, with this dataset may improve the model's accuracy. These models can capture complex linguistic nuances that may aid in truthfulness classification, particularly across varied cultural contexts.

- **Incorporating Multi-Modal Features:** Integrating additional data modalities, such as images or audio, could enrich the model's context-awareness. This would allow the model to capture non-textual clues relevant to the truthfulness of statements, offering a more holistic classification approach.

- **Enhanced Feature Engineering:** Further feature engineering could improve model performance, including exploring new linguistic features or sentiment-based markers. Automated feature selection techniques may help identify the most relevant features, minimizing noise and improving classification accuracy.

- **Feedback-Driven Model Refinement:** Establishing a feedback loop with human evaluators, such as fact-checkers, could help refine the model's predictions over time. Human feedback could guide adjustments for complex or ambiguous cases, leading to improved adaptability and reliability.

In summary, expanding the dataset, fine-tuning with larger models, conducting region-specific evaluations, and incorporating multi-modal data could enhance the model's performance and generalizability. These improvements aim to create a robust model suitable for global applications, reliably assessing truthfulness in diverse contexts.

References

Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings of the First Workshop on Fact Extraction and Verification (FEVER)*, pages 85–90, Brussels, Belgium. Association for Computational Linguistics.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating fact checking explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7352–7364, Online. Association for Computational Linguistics.

William Ferreira and Andreas Vlachos. 2016. Emergent: A novel data-set for stance classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1163–1168, San Diego, California. Association for Computational Linguistics.

Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J. Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management (CIKM)*. ACM.

Bibek Upadhayay and Vahid Behzadan. 2020. Sentimental liar: Extended corpus and deep learning models for fake claim classification. In *Proceedings of the 2020 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 1–6. IEEE.