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)

CSE Department, IIT Bombay

Advisor: Dr. Pushpak Bhattacharyya, Dr. Rudra Murthy

Motivation

- In today's digital age, widespread misinformation is a significant challenge People uses half-truth often in public to mislead people due to the fact that the information is correct.
- No information about the topic being presented, I’m inclined to believe what I hear, especially if the speaker occupies a position of trust and responsibility.

  Example: India won chess olympiad because of Gukesh Dommaraju

- When a speaker presents a topic with which I am quite familiar, it’s much easier to discern when he or she is presenting only half-truths. In that case, I know that the speaker is omitting certain details

  Example: "NLP models like GPT-4 generate human-like text indistinguishable from what a person would write."**Omission:** These models lack actual understanding or intentionality, sometimes generating false or nonsensical information confidently.

Half Truth

- A statement that is only partly true, especially one intended to deceive, evade blame, or the like … a statement that fails to divulge the whole truth presenting only partial information
- Omitting key details or context that could provide a more comprehensive understanding of a topic or situation.
- Half truth primary motive is to deceive or misrepresent the truth.
  - Example: "You should not trust Peter with your children. I once saw him smack a child with his open hand." *In this example the statement could be true, but Peter may have slapped the child on the back because he was choking.*
  - Example: "I'm a really good driver. In the past thirty years, I've gotten only four speeding tickets." *Statement may be true, but is deceptive if speaker started driving a week ago.*

Half truth Examples

- **"Ashwatthama was killed during the Kurukshetra war."**
  - The statement omits the critical detail that the "Ashwatthama" being referred to was actually an elephant named Ashwatthama, not Dronacharya's son. The Pandavas used this partial truth to strategically deceive Drona.
  -
- **"The Biden administration did keep Trump tariffs in place."**
  - This statement is partially true because it is accurate that some tariffs were kept in place, but it is incomplete. The administration did modify and adjust some tariffs, but it also removed others.
  - Evidence: According to the U.S. Trade Representative, the administration has kept some tariffs in place, such as the 25% tariff on steel and aluminum imports, while removing others, such as the 10% tariff on solar panels and washing machines.

Literature Review

- Numerous fact-checking platforms, such as PolitiFact, FactCheck.org, and Snopes, have emerged to combat misinformation by employing rigorous methodologies and have employed a dedicated team of journalists and researchers who thoroughly investigate claims
- In the domain of half-truth detection by Alhindi et al. (2018) has introduced the LIAR-PLUS dataset and involved training the LSTM model to classify statements.
- Apart from half-truth works have been done in detecting fake news and exaggerated news. Wright and Augenstein (2021) supervised learning approach to detect exaggerated claims

Politifact website Truth O Meter

- TRUE – The statement is accurate and there's nothing significant missing.
- MOSTLY TRUE – The statement is accurate but needs clarification or additional information.
- HALF TRUE – The statement is partially accurate but leaves out important details or takes things out of context.
- MOSTLY FALSE – The statement contains an element of truth but ignores critical facts that would give a different impression.
- FALSE – The statement is not accurate and makes a ridiculous claim.

Class Half true and mostly false almost have same definition.

Problem Statement

- **Task:** Given Claim and Evidence detect whether the claim is half true, true, false, mostly-true, mostly-false. Also creating a dataset which can be used to detect whether a statement is half true, true, false, mostly-true, mostly-false

- **Input**: Claim C and Evidence E
- **Output**: One of the five category (true, half-true, false, mostly-true, mostly-false)

Generating Half truth statements

1. Omission: *"Our new product has received rave reviews!"* **Deception**: Omits the fact that only a few reviews were positive, while most were mixed or negative.

2. Exaggeration: *"This is the most revolutionary product in the entire tech industry!"* **Deception**: Overstates the product's significance, implying it has far more impact than it does.

3. Understatement: *"We've made a few small updates to our product."* **Deception**: Downplays major changes, such as a complete redesign or significant feature upgrades, making the improvements seem trivial.

4. Alteration of Facts: *"Our product was named the best in the industry by TechWorld last month!"* **Deception**: Alters the fact, as the product may have been featured in TechWorld, but not necessarily named the best.

5. Over-Representation of Numbers: *"Thousands of users have switched to our product!"* **Deception**: Inflates the number of users switching to their product, when in reality, only a few hundred have done so.

6. Generalization *"Our product has improved user satisfaction across all areas."* **Deception**: Vague and general, this statement doesn't specify which areas were improved, making it sound broader than it actually is.

7. Context Manipulation: *"Our product is 50% faster than the previous version."* **Deception**: Fails to mention that this speed increase only applies to certain specific tasks, not the overall performance of the product.

8. Ambiguity: *"Our product provides unmatched performance."* **Deception**: Vague, with no clear definition of what "unmatched" means or in what context the product performs better.

9. Quantifier Shift: *"Many users prefer our product."* **Deception**: Shifts from a specific quantifier ("most") to a less precise one ("many"), making the endorsement seem larger than it may actually be.

10. Selective Comparison: *"Our product is 20% cheaper than the leading competitor."* **Deception**: This only compares price, omitting that the competitor's product may have more features, better performance, or higher quality.

11. False Equivalence: *"Choosing our product is like choosing the latest smartphone—it's a necessity for everyone."* **Deception**: Equates their product to a smartphone, a much more universally needed item, making the comparison inappropriate.

12. Emotional Appeal: *"If you care about your family's safety, you need our product!"* **Deception**: Plays on emotions (family safety) to obscure the actual merits of the product, making it seem indispensable without providing factual justification.

Omission of Key Information

Leaving out important details can turn a true statement into a half-truth.

Truth Statement : "The company has had record profits this quarter due to the sale of a major asset."

Half truth : "The company has had record profits this quarter."

Explanation : This half-truth omits the critical detail that the profits were due to the sale of a major asset, not from the core business operations. By leaving out this information, the statement gives the misleading impression that the company's ongoing business is performing exceptionally well.

Truth Statement: "Organic food is free from synthetic pesticides and fertilizers, but it is not always more nutritious than conventionally grown food."

Half-Truth: "Organic food is healthier because it is free from synthetic chemicals."
Explanation: This half-truth omits the detail that organic food is not always more nutritious. By focusing only on the absence of synthetic chemicals, the statement misleadingly implies that organic food is always a healthier choice in terms of nutrition.

Existential to Universal Quantifier

Using vague or unclear language can obscure the full truth.

* Example: "Our product is the best on the market." (This is subjective and can be misleading without clear criteria for "best.")

Truth "She was tired; however, she only managed to finish some of her homework"

Half truth- "She was tired; however, she only managed to finish her homework"

1. Some

- **Truth Statement**: "Some people think the policy is beneficial."
- **Half-Truth**: "People think the policy is beneficial."
    - **Explanation**: The half-truth omits the qualifier "some," which is crucial for understanding that not everyone shares this opinion. It misleadingly implies that all people think the policy is beneficial.

2. Many

- **Truth Statement**: "Many students agree with the proposal."
- **Half-Truth**: "Students agree with the proposal."
    - **Explanation**: The half-truth leaves out "many," which indicates that the agreement is not universal. This makes it seem like agreement is more widespread than it actually is.

3. Few

- **Truth Statement**: "Few people attended the meeting."
- **Half-Truth**: "People attended the meeting." **Explanation**: The half-truth omits the word "few," which clarifies that attendance was low. This makes it appear as though attendance was more significant than it actually was.

4. Several

- **Truth Statement**: "Several issues were raised during the discussion."
- **Half-Truth**: "Issues were raised during the discussion."
    - **Explanation**: The half-truth omits "several," which gives the impression that the number of issues raised could be minimal or unspecified, rather than indicating a moderate number.

5. Often

- **Truth Statement**: "The machine often needs repairs."
- **Half-Truth**: "The machine needs repairs."
    - **Explanation**: The half-truth omits "often," which specifies the frequency of repairs. This could mislead someone into thinking repairs are less frequent than they are.

6. Occasionally

- **Truth Statement**: "Occasionally, the system crashes."
- **Half-Truth**: "The system crashes."
    - **Explanation**: By omitting "occasionally," the half-truth makes it seem like system crashes are more frequent or constant than they are.

7. Various

- **Truth Statement**: "The company offers various services."
- **Half-Truth**: "The company offers services."
    - **Explanation**: By omitting "various," the half-truth suggests the range of services might not be as diverse as it actually is.
    -

8. Approximately

- **Truth Statement**: "The project will be completed in approximately two weeks."
- **Half-Truth**: "The project will be completed in two weeks."
    - **Explanation**: The half-truth omits "approximately," which suggests a more precise completion time than the estimate actually provides.

9. All

- **Truth Statement**: "All team members agreed on the new policy."
- **Half-Truth**: "Team members agreed on the new policy."
    - **Explanation**: Omitting "all" makes it unclear whether everyone agreed or only a subset did, potentially misleading the audience about the level of consensus.

10. Any

- **Truth Statement**: "Any employee can submit feedback on the new initiative."
- **Half-Truth**: "Employees can submit feedback on the new initiative."
    - **Explanation**: The half-truth omits "any," which suggests that not all employees might be encouraged or allowed to submit feedback.
    -

11. Most

- **Truth Statement**: "Most customers were satisfied with the service."
- **Half-Truth**: "Customers were satisfied with the service."
    - **Explanation**: Omitting "most" can mislead by implying universal satisfaction, rather than a majority.

Ambiguous Phrases

1. **A lot**
   - **Truth Statement**: "She has a lot of experience."
   - **Half-Truth**: "She has experience."
     - **Explanation**: The half-truth omits "a lot," which downplays the extent of her experience.
2. **In most cases**
   - **Truth Statement**: "In most cases, the procedure is effective."
   - **Half-Truth**: "The procedure is effective."
     - **Explanation**: By omitting "in most cases," the half-truth implies that the procedure is universally effective, which might not be true.
3. **Generally**
   - **Truth Statement**: "The system generally performs well."
   - **Half-Truth**: "The system performs well."
     - **Explanation**: The half-truth omits "generally," which suggests the performance is consistently good rather than a common trend.
4. **On average**
   - **Truth Statement**: "On average, the device lasts for several years."
   - **Half-Truth**: "The device lasts for several years."
     - **Explanation**: The half-truth omits "on average," which could mislead by suggesting all devices last as long as the average.
5. **To some extent**
   - **Truth Statement**: "The policy benefits employees to some extent."
   - **Half-Truth**: "The policy benefits employees."
     - **Explanation**: The half-truth omits "to some extent," which could imply that the policy benefits employees more comprehensively than it actually does.

Context Manipulation

Presenting information out of context can mislead, even if the facts are accurate.

Truth Statement: "The politician voted against a bill that included tax cuts because it also had provisions that would harm public education funding."

Half-Truth: “The politician voted against a bill to lower taxes."

Explanation: This half-truth omits the important detail that the bill also included harmful provisions. Without this context, it misleadingly suggests that the politician is against lowering taxes in general, rather than opposing the specific provisions of the bill.

Truth Statement: "The politician voted for a new environmental policy that includes stricter regulations."

Context Manipulation: "The politician voted for more regulations."

Here, the context manipulation involves stripping away specific details (environmental policy and stricter regulations) to create a misleading impression about the nature of the regulations or the politician’s stance.

Selective Presentation

Highlighting certain facts while ignoring others that are equally relevant.

Full Truth: "Crime rates have dropped in the last year, but they are still higher than they were five years ago."

Half-Truth: "Crime rates have dropped in the last year."

Explanation:

This half-truth omits the longer-term context that crime rates are still higher than they were five years ago. This selective presentation makes it seem as though there has been a significant and sustained improvement in safety, which is misleading.

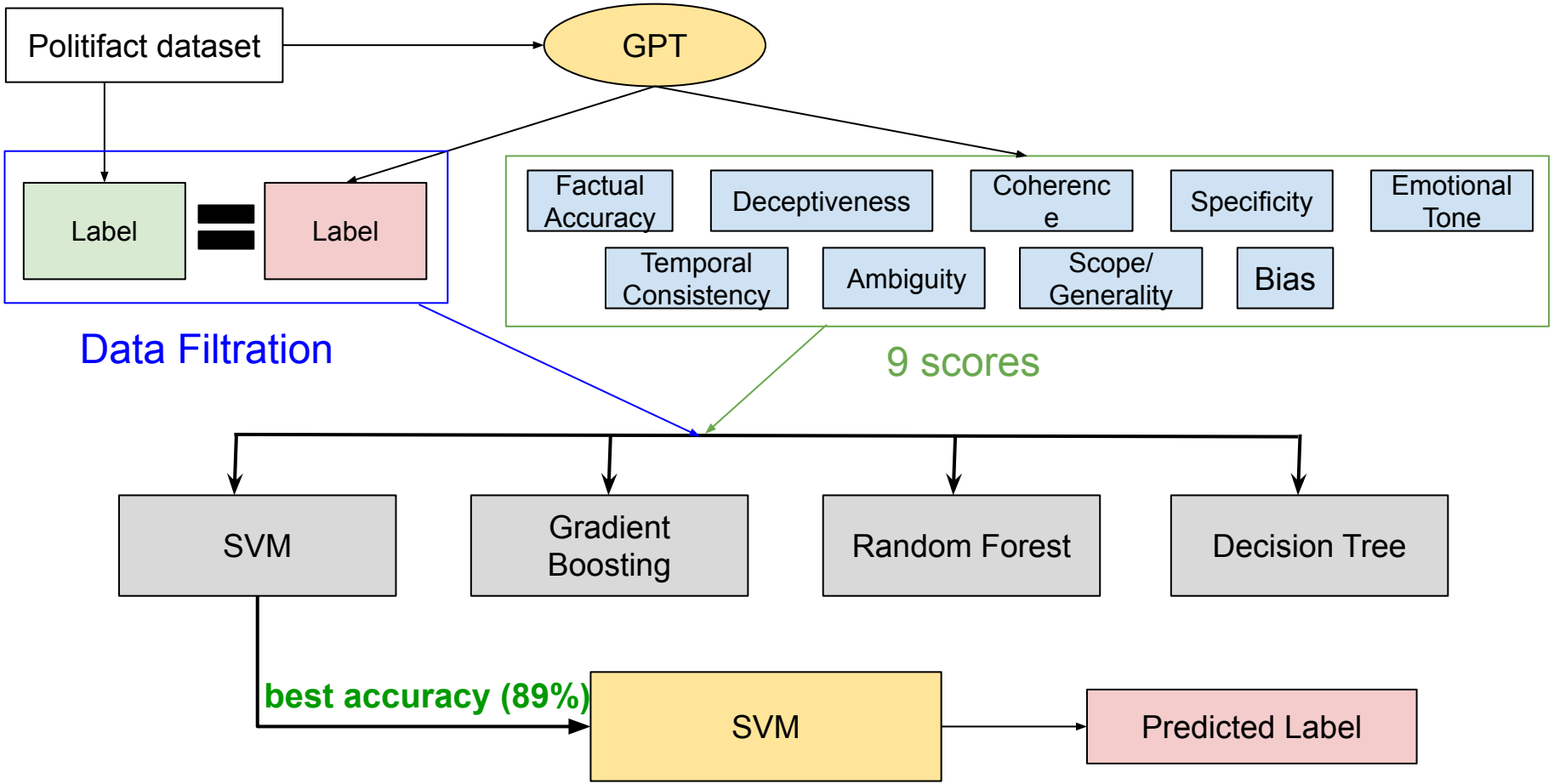
Full Truth: "The politician voted in favor of the new environmental regulations and against some tax cuts."

Half-Truth: "The politician supported the new environmental regulations." (Highlighting only the positive action while omitting their opposition to tax cuts, which might also be relevant to understanding their overall stance.)

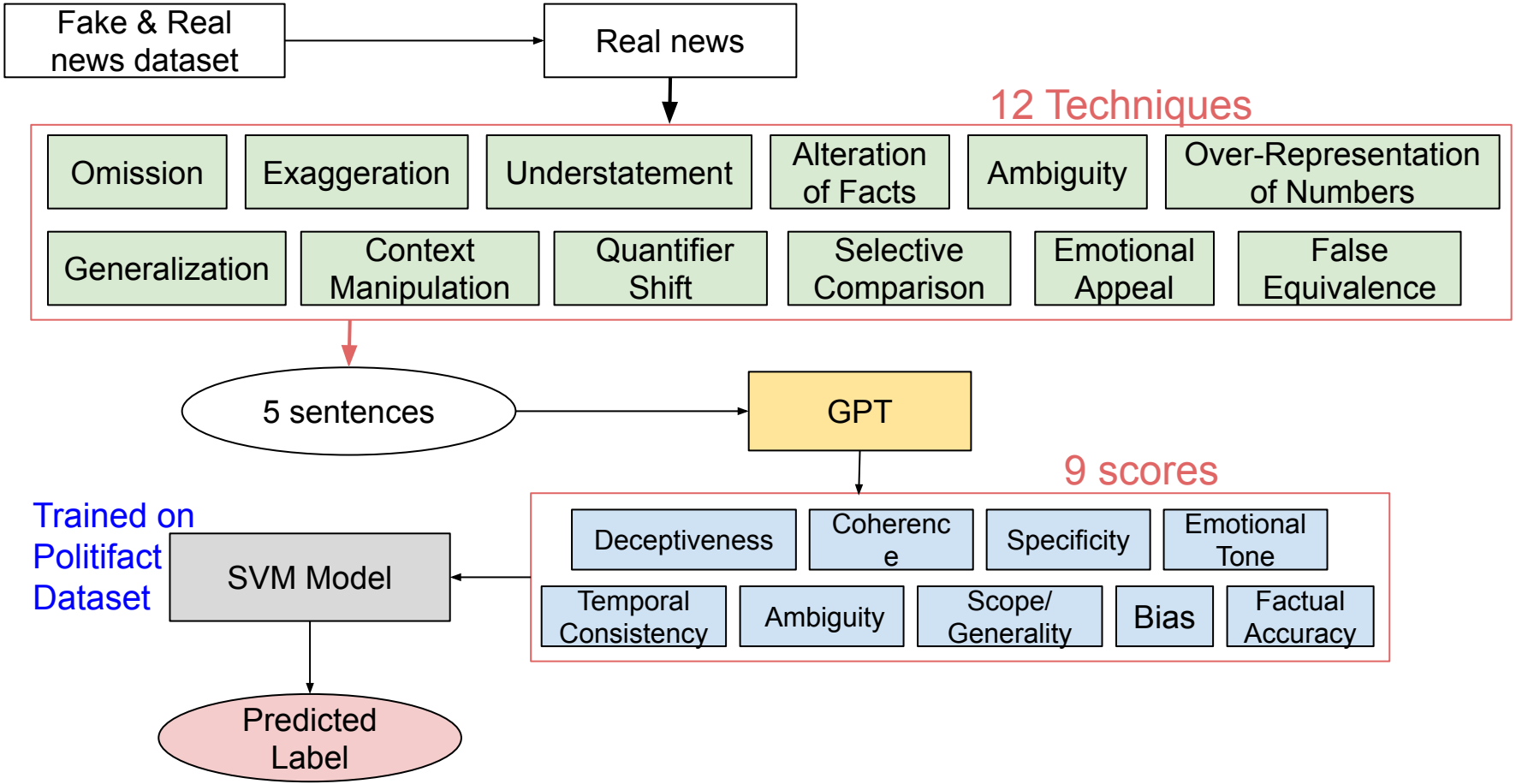
Scoring Criteria for Deception Statements

- **Factual Accuracy (0-1):**
  How closely the statement matches the facts.
  Score: 0 = Inaccurate, 1 = Accurate.
- **Deceptiveness (0-1):**
  Measures the potential to mislead or distort.
  Score: 0 = Not deceptive, 1 = Highly deceptive.
- **Coherence (0-1):**
  Logical flow and clarity of the statement.
  Score: 0 = Incoherent, 1 = Clear.
- **Specificity (0-1):**
  Level of detail provided.
  Score: 0 = Vague, 1 = Specific.
- **Emotional Tone (0-1):**
  Whether the tone evokes emotions.
  Score: 0 = Neutral, 1 = Emotional.
- **Bias (0-1):**
  Presence of bias or unbalanced perspective.
  Score: 0 = Neutral, 1 = Biased.
- **Scope (0-1):**
  How broad or specific the statement is.
  Score: 0 = Specific, 1 = General.
- **Temporal Consistency (0-1):**
  Accuracy of the statement in relation to time.
  Score: 0 = Misleading, 1 = Accurate.
- **Context/Ambiguity (0-1):**
  Whether key context or clarity is missing.
  Score: 0 = Ambiguous, 1 = Clear.

Pipeline for Model Detecting labels



Pipeline for Dataset Creation



Task: Generate deceptive statements based on the following original paragraph by applying one or more of the following techniques mentioned below You may combine categories to create more nuanced or complex misrepresentations.

Techniques You Can Apply:

Paraphrasing: Slightly modify the original statement while keeping its essence, introducing subtle misrepresentation.

Text Perturbations: Modify words (e.g., using negation or synonyms) to alter the meaning.

Adversarial Attacks: Generate deceptive versions by altering key phrases or the sentiment of the original text.

Omission: Leave out key facts that change the context of the original paragraph.

Exaggeration: Overstate certain points to make the situation seem more severe.

Understatement: Downplay the importance of key points.

Alteration of Facts: Change specific details such as dates, numbers, or institutions.

Over-Representation of Numbers: Inflate or distort numerical data.

Generalization: Make the statement more vague to obscure critical details.

Context Manipulation: Change the surrounding context to mislead readers.

Ambiguity: Use vague or unclear language.

Quantifier Shift: Alter quantifiers to distort the magnitude of an event.

Selective Comparison: Only mention favorable facts for a biased representation.

False Equivalence: Compare this situation to an unrelated one to create a misleading connection.

Misleading Cause and Effect: Imply that one event caused another without proof.

Emotional Appeal: Use sensational language to distract from facts.

Experiment 1: Initial Model (PolitiFact Data)

- **Dataset**: PolitiFact dataset with 250 instances (50 per class: *True, Half-True, Mostly-True, Mostly-False, False*). Each instance included a claim and evidence.
- **Features**: Deceptiveness, Factual Accuracy, and Coherence scores generated using GPT-4 prompts
- **Gradient Boosting** performed best, achieving 68% accuracy. It adapted well to noisy or less informative features by learning sequentially from mistakes. Its iterative nature allowed it to optimize performance effectively, especially when working with a small dataset containing complex patterns.

Comparison between different models

Model	Accuracy	Macro F1 Score	Macro Recall	Macro Precision
Decision Tree	0.58	0.58	0.59	0.60
SVM	0.58	0.56	0.56	0.57
Random Forest	0.64	0.63	0.64	0.63
Gradient Boosting	0.68	0.66	0.69	0.70

Decision Tree

Accuracy: 0.58

Classification Report

Class	Precision	Recall	F1-score	Support
0 Half-true	0.36	0.71	0.48	7
1 Mostly-false	0.67	0.60	0.63	10
2 Mostly-true	0.44	0.36	0.40	11
3 False	0.89	0.62	0.73	13
4 True	0.67	0.67	0.67	9
Macro Avg	0.60	0.59	0.58	50
Weighted Avg	0.64	0.58	0.59	50

SVM

Accuracy: 0.58

Classification Report

Class	Precision	Recall	F1-score	Support
0 Half-true	0.33	0.43	0.38	7
1 Mostly-false	0.62	0.50	0.56	10
2 Mostly-true	0.45	0.45	0.45	11
3 False	0.83	0.77	0.80	13
4 True	0.60	0.67	0.63	9
Macro Avg	0.57	0.56	0.56	50
Weighted Avg	0.60	0.58	0.59	50

Random Forest

Accuracy: 0.64
Classification Report

Class	Precision	Recall	F1-score	Support
0 Half-true	0.50	0.71	0.59	7
1 Mostly-false	0.62	0.50	0.56	10
2 Mostly-true	0.56	0.45	0.50	11
3 False	0.83	0.77	0.80	13
4 True	0.64	0.78	0.70	9
Macro Avg	0.63	0.64	0.63	50
Weighted Avg	0.65	0.64	0.64	50

Gradient Boosting Classifier

Accuracy: 0.68

Classification Report

Class	Precision	Recall	F1-score	Support
0 Half-true	0.45	0.71	0.56	7
1 Mostly-false	0.71	0.50	0.59	10
2 Mostly-true	0.83	0.45	0.59	11
3 False	0.83	0.77	0.80	13
4 True	0.64	1.00	0.78	9
Macro Avg	0.70	0.69	0.66	50
Weighted Avg	0.72	0.68	0.67	50

Experiment 2: Feature Engineering to Improve Model Accuracy

- Dropped: The Coherence feature based on low contribution from Experiment 1 and new features such as Specificity, Emotional Tone, Scope/Generality, Temporal Consistency, and Out of Context or Ambiguity were added.
- All models showed **consistent accuracy (70-74%)**, suggesting the added features enhanced feature richness
- **SVM** outperformed other models with the highest **Macro F1-Score (0.74)**, demonstrating better precision-recall balance across all classes.
- The higher performance of SVM indicates that the model benefited from the refined feature set, likely due to its ability to find hyperplane margins effectively with limited noise.
- Classes 3 and 4 (True, False) consistently achieved better scores, showing that more explicit lies or truths are easier to classify than ambiguous or mixed claims.

Comparison between different models

Model	Accuracy	Macro F1-Score	Macro Recall	Macro Precision
Decision Tree	0.70	0.70	0.70	0.70
Random Forest	0.70	0.70	0.70	0.70
Gradient Boosting	0.70	0.70	0.70	0.70
SVM	0.74	0.74	0.74	0.75

Decision Tree (Removed the Coherence column)

Accuracy: 0.7

Classification Report:

	precision	recall	f1-score	support
0	0.60	0.60	0.60	10
1	0.64	0.70	0.67	10
2	0.67	0.60	0.63	10
3	0.80	0.80	0.80	10
4	0.80	0.80	0.80	10
accuracy			0.70	50
macro avg	0.70	0.70	0.70	50
weighted avg	0.70	0.70	0.70	50

Random Forest (Removed the Coherence column)

Accuracy: 0.7

Classification Report:

	precision	recall	f1-score	support
0	0.60	0.60	0.60	10
1	0.60	0.60	0.60	10
2	0.75	0.60	0.67	10
3	0.73	0.80	0.76	10
4	0.82	0.90	0.86	10
accuracy			0.70	50
macro avg	0.70	0.70	0.70	50
weighted avg	0.70	0.70	0.70	50

Gradient Boosting Classifier(Removed the Coherence column)

Accuracy: 0.7

Classification Report:

	precision	recall	f1-score	support
0	0.60	0.60	0.60	10
1	0.60	0.60	0.60	10
2	0.75	0.60	0.67	10
3	0.73	0.80	0.76	10
4	0.82	0.90	0.86	10
accuracy			0.70	50
macro avg	0.70	0.70	0.70	50
weighted avg	0.70	0.70	0.70	50

SVM (Removed the Coherence column)

Accuracy: 0.74

Classification Report

Class	Precision	Recall	F1-score	Support
0	0.62	0.80	0.70	10
1	0.75	0.60	0.67	10
2	0.75	0.60	0.67	10
3	0.80	0.80	0.80	10
4	0.82	0.90	0.86	10
Macro Avg	0.75	0.74	0.74	50
Weighted Avg	0.75	0.74	0.74	50

Experiment 3: Score Generation with GPT-4 Mini

- Generated: 250 synthetic instances (50 per class) using GPT-4 Mini Feature Set: Same as Experiment 2.
- **SVM** achieved the highest **accuracy (72%)** and **macro F1-Score (0.71)**, demonstrating its ability to generalize across the dataset.
- **Gradient Boosting** performed well with **70% accuracy** and balanced metrics, while **Random Forest,** and **Decision Tree** had slightly lower scores.
- Class **4 (False)** consistently scored the highest across all models, suggesting that the data captures falsehoods clearly and models can distinguish them effectively.
- Lower performance for Class **1 (Mostly-True)** across some models highlights challenges in identifying less explicit statements.

Comparison between different models

Model	Accuracy	Macro F1-Score	Macro Recall	Macro Precision
SVM	0.72	0.71	0.72	0.75
Random Forest	0.68	0.66	0.68	0.73
Decision Tree	0.66	0.63	0.66	0.72
Gradient Boosting Classifier	0.70	0.70	0.70	0.70

SVM (GPT-4o mini)

Accuracy: 0.72

Classification Report:

	precision	recall	f1-score	support
0	0.58	0.70	0.64	10
1	1.00	0.50	0.67	10
2	0.67	0.60	0.63	10
3	0.75	0.90	0.82	10
4	0.75	0.90	0.82	10
accuracy			0.72	50
macro avg	0.75	0.72	0.71	50
weighted avg	0.75	0.72	0.71	50

Random Forest

Accuracy: 0.68

Classification Report:

	precision	recall	f1-score	support
0	0.58	0.70	0.64	10
1	1.00	0.30	0.46	10
2	0.67	0.60	0.63	10
3	0.64	0.90	0.75	10
4	0.75	0.90	0.82	10
accuracy			0.68	50
macro avg	0.73	0.68	0.66	50
weighted avg	0.73	0.68	0.66	50

Decision Tree

Accuracy: 0.66

Classification Report:

	precision	recall	f1-score	support
0	0.58	0.70	0.64	10
1	1.00	0.20	0.33	10
2	0.67	0.60	0.63	10
3	0.60	0.90	0.72	10
4	0.75	0.90	0.82	10
accuracy			0.66	50
macro avg	0.72	0.66	0.63	50
weighted avg	0.72	0.66	0.63	50

Light Gradient Boosting Machine

Accuracy: 0.68

Classification Report:

	precision	recall	f1-score	support
0	0.43	0.86	0.57	7
1	0.80	0.40	0.53	10
2	0.62	0.45	0.53	11
3	0.92	0.92	0.92	13
4	0.70	0.78	0.74	9
accuracy			0.68	50
macro avg	0.70	0.68	0.66	50
weighted avg	0.72	0.68	0.68	50

Gradient Boosting Classifier

Accuracy: 0.7

Classification Report:

	precision	recall	f1-score	support
0	0.60	0.60	0.60	10
1	0.60	0.60	0.60	10
2	0.75	0.60	0.67	10
3	0.73	0.80	0.76	10
4	0.82	0.90	0.86	10
accuracy			0.70	50
macro avg	0.70	0.70	0.70	50
weighted avg	0.70	0.70	0.70	50

Experiment 4: Experiment 4: Data Filtering

- Processed data generated from chatgpt.(200 Instances)
- To increase confidence in the scores generated by ChatGPT, only those data points where the predicted label from ChatGPT accurately matched the ground truth label were retained for further analysis
- **SVM**: Accuracy: **0.85,** with high recall for label 0 and f1-scores above 0.80 for most classes.
- Filtering based on GPT-4 Mini's accurate predictions improved model accuracy and highlighted the importance of consistent pre-processing.

Class	Precision	Recall	F1-Score	Support
Half-true	0.67	1.00	0.80	8
Mostly-false	1.00	0.50	0.67	8
Mostly-true	0.86	0.75	0.80	8
False	1.00	1.00	1.00	8
True	0.89	1.00	0.94	8
Overall Accuracy			0.85	40
Macro Average	0.88	0.85	0.84	40
Weighted Average	0.88	0.85	0.84	40

SVM using gpt-4o-mini(200 instances)

Accuracy: 0.85

Classification Report:

	precision	recall	f1-score	support
0	0.67	1.00	0.80	8
1	1.00	0.50	0.67	8
2	0.86	0.75	0.80	8
3	1.00	1.00	1.00	8
4	0.89	1.00	0.94	8
accuracy			0.85	40
macro avg	0.88	0.85	0.84	40
weighted avg	0.88	0.85	0.84	40

Experiment 5: Enlarging Dataset and Feature Optimization

- Setup: Increased dataset size to 400 which after preprocessing remained 382 and then identified optimal feature combinations for SVM.
- **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) boosted accuracy.

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
Overall Accuracy			0.90	77
Macro Average	0.91	0.88	0.88	77
Weighted Average	0.91	0.90	0.89	77

Analysis (Best Model)

- **Class 1 (Mostly False)**: While precision was perfect (1.00), recall dropped to **0.54**, suggesting the model struggled to identify all instances of this class.
- **Classes 2 (Mostly True), 3 (True), and 4 (False)**: Achieved excellent recall (1.00), reflecting the model's capability to consistently classify these labels.
- **Class 0 (Half True)**: Balanced performance with **0.88** for both precision and recall, indicating a reliable prediction rate.

Best Parameters Selection Using GridSearchCV: Explores a **hyperparameter grid** for SVM (e.g., `C`, `kernel`, `gamma`) Ensures exhaustive search for the best configuration.

Cross-Validation for Reliable Performance

- **Stratified K-Fold Cross-Validation**:
  - Splits data into **K folds** (5 folds in this case) with preserved class distribution across folds.
  - Ensures balanced representation of all classes in each fold, especially important for **imbalanced datasets**.
  - Model is trained on K−1K-1K−1 folds and tested on the remaining fold. This process is repeated KKK times, providing robust evaluation.

Overall Metrics:

- **Macro Avg Precision (0.91)**: Reflects the model's ability to maintain high precision across all classes.
- **Macro Avg Recall (0.88)**: Indicates slightly lower sensitivity for specific classes (e.g., Class 1).
- **Weighted Avg F1-Score (0.89)**: Demonstrates a balanced performance, considering class distributions.

Dataset Impact:

- Increasing the dataset size to 382 instances enhanced model training, leading to better generalization and accuracy.

Gradient boosting

Accuracy: 0.8701298701298701

Classification Report:

	precision	recall	f1-score	support
0	0.87	0.81	0.84	16
1	0.88	0.54	0.67	13
2	0.82	0.93	0.88	15
3	0.84	1.00	0.91	16
4	0.94	1.00	0.97	17
accuracy			0.87	77
macro avg	0.87	0.86	0.85	77
weighted avg	0.87	0.87	0.86	77

Experiment 6: Mistral

- Setup: Explore model performance using Mistral-generated scores.
- **SVM** achieved the highest **accuracy (0.78)** and **macro F1-score (0.78)**, demonstrating its robustness in handling the generated feature set.

Model	Accuracy	Macro F1-Score	Macro Recall	Macro Precision
SVM	0.78	0.78	0.78	0.79
Random Forest	0.68	0.67	0.68	0.68
Decision Tree	0.70	0.70	0.70	0.78
Gradient Boosting Classifier	0.70	0.70	0.70	0.70

SVM (Mistral)

Accuracy: 0.78

Classification Report:

	precision	recall	f1-score	support
0	0.75	0.60	0.67	10
1	0.89	0.80	0.84	10
2	0.88	0.70	0.78	10
3	0.77	1.00	0.87	10
4	0.67	0.80	0.73	10
accuracy			0.78	50
macro avg	0.79	0.78	0.78	50
weighted avg	0.79	0.78	0.78	50

Random Forest(Mistral)

Accuracy: 0.68

Classification Report:

	precision	recall	f1-score	support
0	0.45	0.50	0.48	10
1	0.86	0.60	0.71	10
2	0.50	0.40	0.44	10
3	0.77	1.00	0.87	10
4	0.82	0.90	0.86	10
accuracy			0.68	50
macro avg	0.68	0.68	0.67	50
weighted avg	0.68	0.68	0.67	50

Decision Tree

Accuracy: 0.7

Classification Report:

	precision	recall	f1-score	support
0	0.50	1.00	0.67	10
1	0.62	0.50	0.56	10
2	1.00	0.50	0.67	10
3	0.88	0.70	0.78	10
4	0.89	0.80	0.84	10
accuracy			0.70	50
macro avg	0.78	0.70	0.70	50
weighted avg	0.78	0.70	0.70	50

Gradient Boosting Classifier

Accuracy: 0.7

Classification Report:

	precision	recall	f1-score	support
0	0.60	0.60	0.60	10
1	0.60	0.60	0.60	10
2	0.75	0.60	0.67	10
3	0.73	0.80	0.76	10
4	0.82	0.90	0.86	10
accuracy			0.70	50
macro avg	0.70	0.70	0.70	50
weighted avg	0.70	0.70	0.70	50

Light Gradient Boosting Machine

Accuracy: 0.68

Classification Report:

	precision	recall	f1-score	support
0	0.43	0.86	0.57	7
1	0.80	0.40	0.53	10
2	0.62	0.45	0.53	11
3	0.92	0.92	0.92	13
4	0.70	0.78	0.74	9
accuracy			0.68	50
macro avg	0.70	0.68	0.66	50
weighted avg	0.72	0.68	0.68	50

Fine Tuning Experiments for Custom data

Experiment 1: PolitiFact 800-Instance Dataset (Train), Full PolitiFact Dataset (Test & Validation)

- **Accuracy:** 39% (moderate improvement).
- **Findings:**
  - **Class Performance:**
    - *False* retained good precision (0.67) and recall (0.60), performing best among all classes.
    - *Mostly-True* and *Half-True* saw minor improvements in recall and f1-scores.
  - **Insights:** Training on PolitiFact data improved performance, highlighting the importance of aligning the training and testing datasets.

Experiment 2: Full PolitiFact Dataset (Train, Test, Validation)

- **Accuracy:** 48% (highest among all experiments).
- **Findings:**
  - **Class Performance:**
    - *True, Mostly-True, and Half-True:* Improved precision and recall (precision ~0.31-0.43).
    - *False:* Maintained high precision and recall (0.71 and 0.72).
    - Overall balance across classes.
  - **Insights:** Full dataset utilization allowed the model to learn more nuanced patterns, improving both precision and recall.

Fine Tuning Experiments for Custom data

Experiment 1: Custom 800-Instance Dataset (Train), Full PolitiFact Dataset (Test & Validation)

- **Accuracy:** 30% (lowest among all experiments).
- **Findings:**
  - **Class Performance:**
    - *False* had relatively high precision (0.82) but low recall (0.31), suggesting it was over-predicted.
    - Other classes suffered from poor recall and precision, particularly *True* (precision: 0.24, recall: 0.19).
  - **Challenges:** The mismatch between training (custom data) and testing (full PolitiFact) created inconsistencies in model generalization.

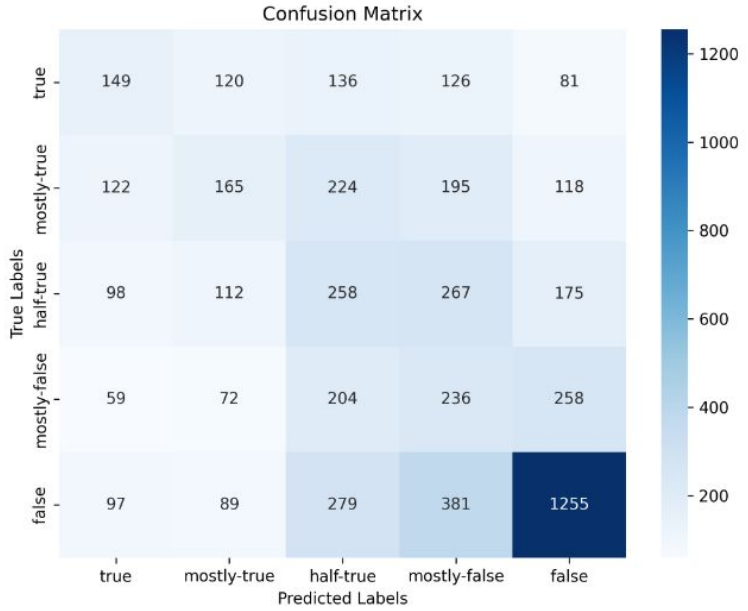
Experiment 4: Custom Dataset (640 Train, 100 Test, 60 Validation)

- **Accuracy:** 38% (similar to Experiment 2).
- **Findings:**
  - **Class Performance:**
    - *Mostly-True* showed the highest recall (0.65), indicating strong performance on this class.
    - Other classes had low precision and recall, especially *True* (precision: 0.50, recall: 0.15).
  - **Insights:** The smaller dataset size limited the model's ability to generalize.

Classification Report: Politifact Data (Train: 900, Test Full, Validation Full)

Classification Report

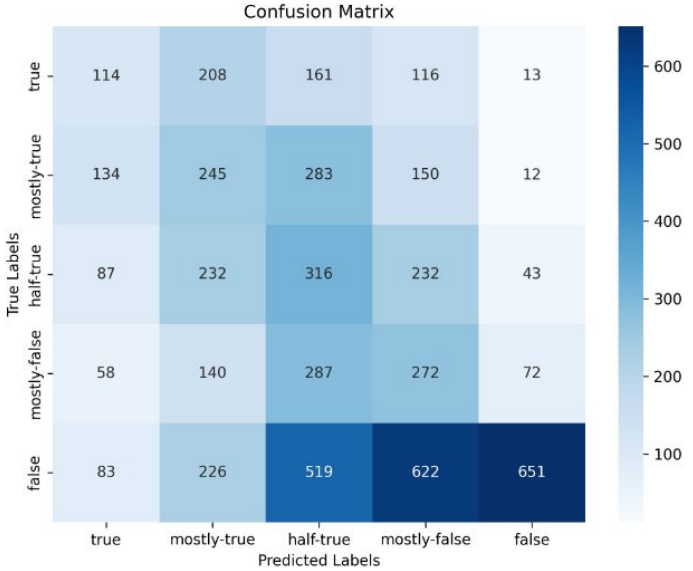
Class	Precision	Recall	F1-Score	Support
True	0.28	0.24	0.26	612
Mostly-True	0.30	0.20	0.24	824
Half-True	0.23	0.28	0.26	910
Mostly-False	0.20	0.28	0.23	829
False	0.67	0.60	0.63	2101
Overall Accuracy			0.39	5276
Macro Average	0.33	0.32	0.32	5276
Weighted Average	0.42	0.39	0.40	5276



Classification Report: Custom Data (Train: 900, Test Full, Validation Full)

Classification Report

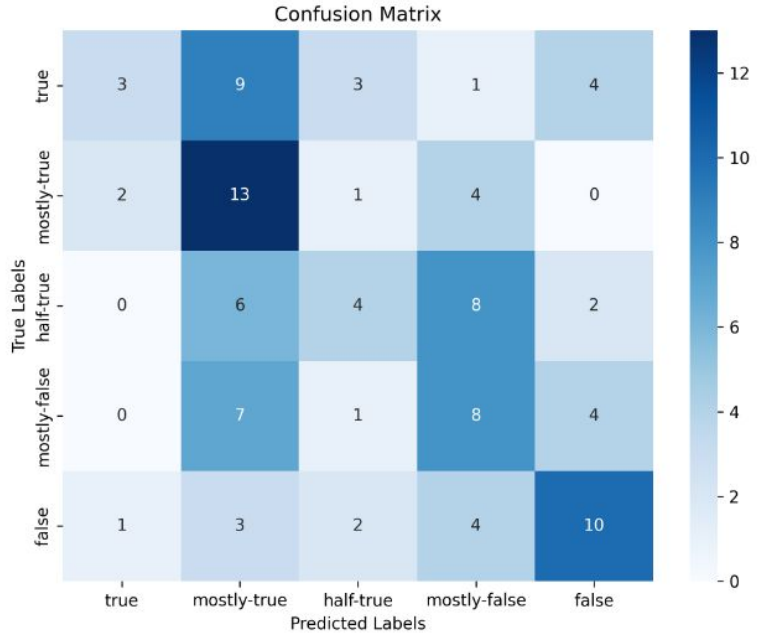
Class	Precision	Recall	F1-Score	Support
True	0.24	0.19	0.21	612
Mostly-True	0.23	0.30	0.26	824
Half-True	0.20	0.35	0.26	910
Mostly-False	0.20	0.33	0.24	829
False	0.82	0.31	0.45	2101
Overall Accuracy			0.30	5276
Macro Average	0.34	0.29	0.28	5276
Weighted Average	0.46	0.30	0.33	5276



Classification Report: Custom Data (Train: 740, Test 100, Validation 60)

Classification Report

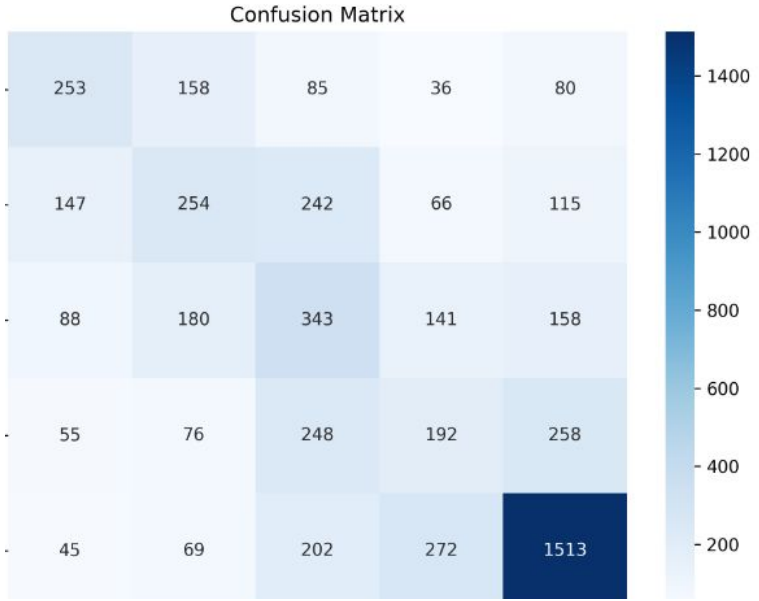
Class	Precision	Recall	F1-Score	Support
True	0.50	0.15	0.23	20
Mostly-True	0.34	0.65	0.45	20
Half-True	0.36	0.20	0.26	20
Mostly-False	0.32	0.40	0.36	20
False	0.50	0.50	0.50	20
Overall Accuracy			0.38	100
Macro Average	0.41	0.38	0.36	100
Weighted Average	0.41	0.38	0.36	100



Performance Metrics on Politifact Full and Test Full Datasets

Classification Report

Class	Precision	Recall	F1-Score	Support
True	0.43	0.41	0.42	612
Mostly-True	0.34	0.31	0.33	824
Half-True	0.31	0.38	0.34	910
Mostly-False	0.27	0.23	0.25	829
False	0.71	0.72	0.72	2101
Overall Accuracy			0.48	5276
Macro Average	0.41	0.41	0.41	5276
Weighted Average	0.48	0.48	0.48	5276



Train 14771
Val 1055
Test 5276

Conclusion

- The importance of half-truth detection lies in its significant societal impact.
- We analyzed how a statement can be transformed into a half-truth.
- A model was developed to classify statements into one of five categories, with the SVM model achieving the best performance at 89% accuracy.
- A custom dataset was created and fine-tuned using XLM-RoBERTa, and its performance was thoroughly analyzed.

Future Work and Limitations

Limitations

- **Dataset Source Mismatch:**
  - Training data: Indian news statements.
  - Testing data: U.S.-centered PolitiFact dataset.
  - Result: Cultural and linguistic mismatches lowered accuracy.

Future Work

1. **Cross-Regional Dataset Expansion**
   - Include data from Indian, U.S., and global news sources for better generalization.
2. **Fine-Tuning Larger LLMs**
   - Use models like GPT-4 or LLaMA for improved accuracy across varied linguistic contexts.
3. **Multi-Modal Integration**
   - Incorporate images, audio, or video to capture non-textual cues for enhanced truthfulness assessment.
4. **Human Feedback Loop**
   - Collaborate with fact-checkers to refine the model.

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.

Dustin Wright and Isabelle Augenstein. 2021. Semi Supervised exaggeration detection of health science press releases. CoRR, abs/2108.13493.

Singamsetty, Sandeep, Nishtha Madaan, Sameep Mehta, Varad Bhatnagar, and Pushpak Bhattacharyya. "" Beware of deception": Detecting Half-Truth and Debunking it through Controlled Claim Editing." *arXiv preprint arXiv:2308.07973* (2023).

https://www.politifact.com/article/2018/feb/12/principles-truth-o-meter-politifacts-methodology-i/

Thank you

EXamples

Half-Truth Statement:

Claim: "We had the greatest economy in the history of our country."

Evidence: Trump frequently repeated this claim throughout the debate but provided no factual evidence to support it.

Reality: Economic data shows that the U.S. economy experienced significant growth during Trump's presidency, but it also experienced the highest inflation rate in over 40 years.

Half-Truth Statement:

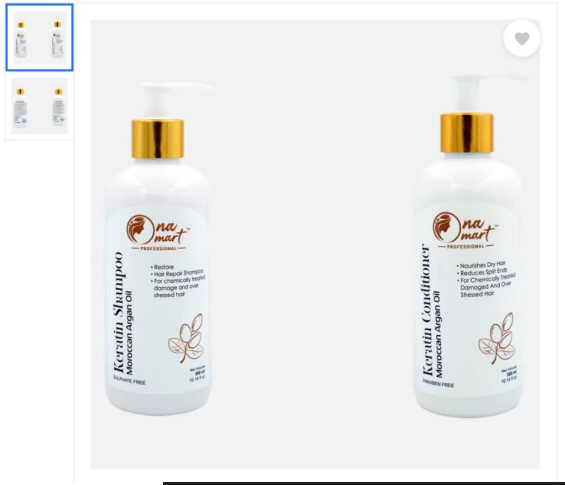
Claim: "The Biden administration kept many of Trump's tariffs in place."

Evidence: While some Trump tariffs were kept in place by the Biden administration, others were immediately lifted or reduced.

Reality: The Biden administration pursued a different trade policy than the Trump administration, focusing on multilateralism and diplomacy.

Half Truth in Cosmetic Products

What do they mean by HD GLOW?



Home › Beauty and G... › Hair Care and... › Hair Care › Hair Oil › ONAMART H... › ONAMART E...

ONAMART ENRICHED WITH MOROCCAN ARGAN OIL Hair Oil (600 ml)

Be the first to Review this product

₹1,795 ~~₹1,798~~ @299.167/100ml

Available offers

- Partner Offer Sign-up for Flipkart Pay Later & get free Times Prime Benefits worth ₹20,000* Kn
- Partner Offer Make a purchase and enjoy a surprise cashback/ coupon that you can redeem la
- Bank Offer Get ₹25* instant discount for the 1st Flipkart Order using Flipkart UPI T&C
- Bank Offer 5% Cashback on Flipkart Axis Bank Card T&C

View 18 more offers

Delivery 400076 Change Check pincode

Delivery by 8 May, Wednesday | ₹80

View Details

Argan oil can cost as much as $300 per liter, making it the world's most expensive edible oil. Made in Morocco. 13 Dec 2023

twitter.com
https://twitter.com › BusinessInsider › status

Business Insider on X: "This is why argan oil is the world's most ...

https://www.flipkart.com/onamart-enriched-moroccan-argan-oil-hair/p/itm2ce317bb2e073
https://twitter.com/BusinessInsider/status/1735030243008672191
https://images.app.goo.gl/JJAJPkoQpNHrmGzMA

SVM(Best)

Best parameters: {'svm__C': 1, 'svm__gamma': 'scale', 'svm__kernel': 'linear'}

Cross-Validation Accuracy Scores: [0.83606557 0.90163934 0.83606557 0.8852459 0.80327869]

Mean Cross-Validation Accuracy: 0.8524590163934427

Standard Deviation of Cross-Validation Accuracy: 0.0359162332790272 8