**1. Problem Statement**

Fake news detectors (AI models that classify news as real or fake) can be **tricked by small, clever changes** in the text. Attackers can modify fake news slightly so the detector **wrongly labels it as real**.

* **How easily can fake news detectors be fooled?**
* **Which models are most vulnerable?**
* **How can we make detectors more robust?**

**2. Solution Approach**

**Step-1 : Generate Adversarial Fake News**

* **Word Swaps** ("false" → "misleading")
* **Typos** ("study" → "studdy")
* **Add Neutral Words** ("actually," "really")

**Example:**

* Original (Fake): *"The vaccine causes dangerous side effects."*
* Modified (Fools AI): *"The vaccine actually causes some incorrect side effects."*

**Step – 2 : Test on Fake News Detectors**

**Models to Test:**

* **Strong Models:** BERT, RoBERTa, DistilBERT
* **Weaker Models (for comparison):** LSTM, Logistic Regression
* **State-of-the-art Model:** Grover (for AI-generated fake news)

**Datasets for Testing:**

* **FakeNewsNet (** [**Link**](https://github.com/KaiDMML/FakeNewsNet/tree/master/dataset) **)**
* **LIAR (** [**Link**](https://www.cs.ucsb.edu/~william/data/liar_dataset.zip) **)**
* **COVID-19 Fake News Dataset**

**3. Solution Summary**

**Adversarial Training (Most Effective Defense) -** *Train the model on both original and attacked fake news*

**How?**

* **Generate Adversarial Examples**
* **Mix Adversarial + Clean Data**
* **Re-Train the Model**

**Evaluation Metrics**

**Mix of Standard Classification Metrics + Adversarial Robustness Metrics**

**1. Standard Classification Metrics**

* **Accuracy -** *Overall correctness*
* **Precision -** *Measures how many "fake" predictions were correct*
* **Recall** *- Measures how many actual fakes were caught*
* **F1 – Score -** *Balances precision and recall*
* **AUC Curve** *- Shows model’s ability to distinguish classes*

**2. Project Specific Metrics ( Adversarial Robustness Metrics )**

* **Attack Success Rate (ASR) -** *% of adversarial examples that fooled the model*
* **Robust Accuracy -** *How many samples were correctly classified*
* **Clean-to-Adversarial Gap -** *Measures performance drop due to attacks*
* **Confidence Drop -** *If attacks reduce model confidence, it’s a sign of instability*

**4. Research Paper Structure**

1. **Title** – *"Fooling the Fact-Checkers: Adversarial Attacks on Fake News Detectors"*
2. **Abstract** – Summary of problem, methods, results.
3. **Introduction** – Why fake news detectors are vulnerable.
4. **Related Work** – Past research on attacks & defenses.
5. **Methodology** – How attacks were generated & tested.
6. **Experiments** – Results on different models.
7. **Discussion** – Why some models fail, real-world risks.
8. **Defenses** – How to improve robustness.
9. **Conclusion** – Key takeaways & future work.