**Adversarial attacks on text classification models**

In this video, we will look into adversarial attacks, explore their implications on text classification models, and discuss how to shield our AI from these threats.

**What are adversarial attacks?**

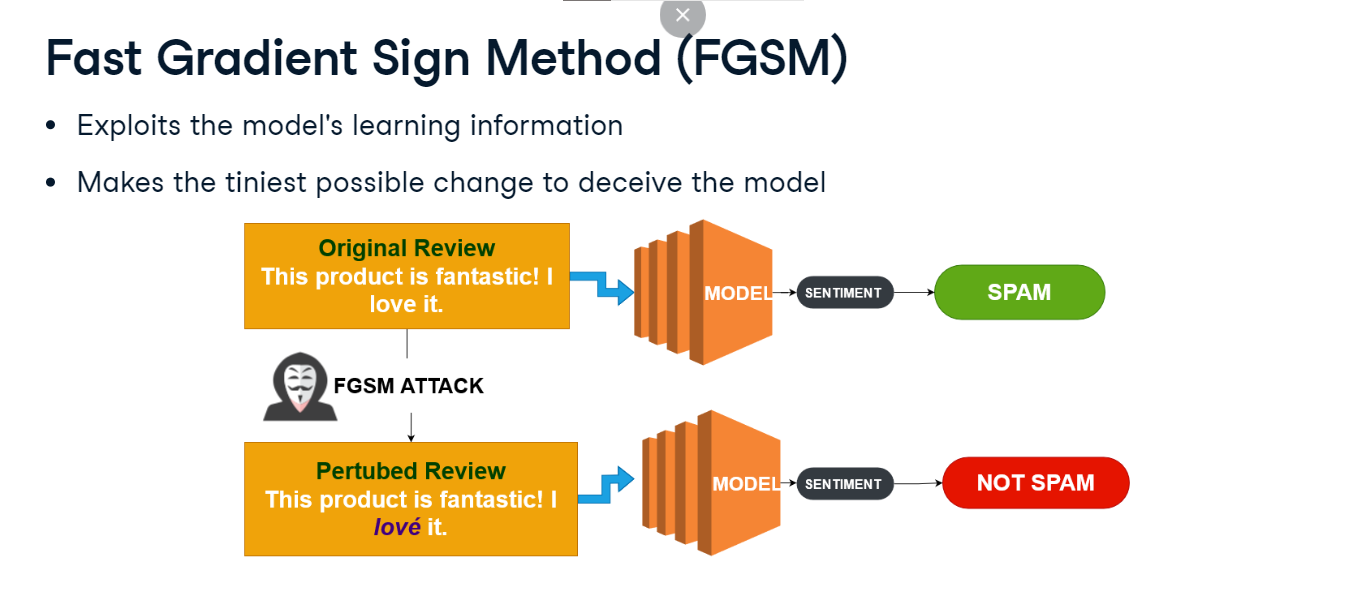
At their core, adversarial attacks are about making crafty tweaks to input data. These aren't just random distortions but calculated malicious changes that can drastically skew the decision-making of an AI model.

**Importance of robustness**

AI's role in interpreting and responding to text is monumental. Consider an AI system tasked with moderating online comments. It must differentiate between healthy dialogue and toxic content. But adversarial attacks can twist this. A well-crafted adversarial input might make benign comments appear harmful or vice versa. Then there's the danger of biased training data: an AI can unintentionally amplify negative stereotypes from biased data. And the perils grow: envision querying an AI chat service about the weather, but receiving misleading information because it was deceived by malicious inputs.

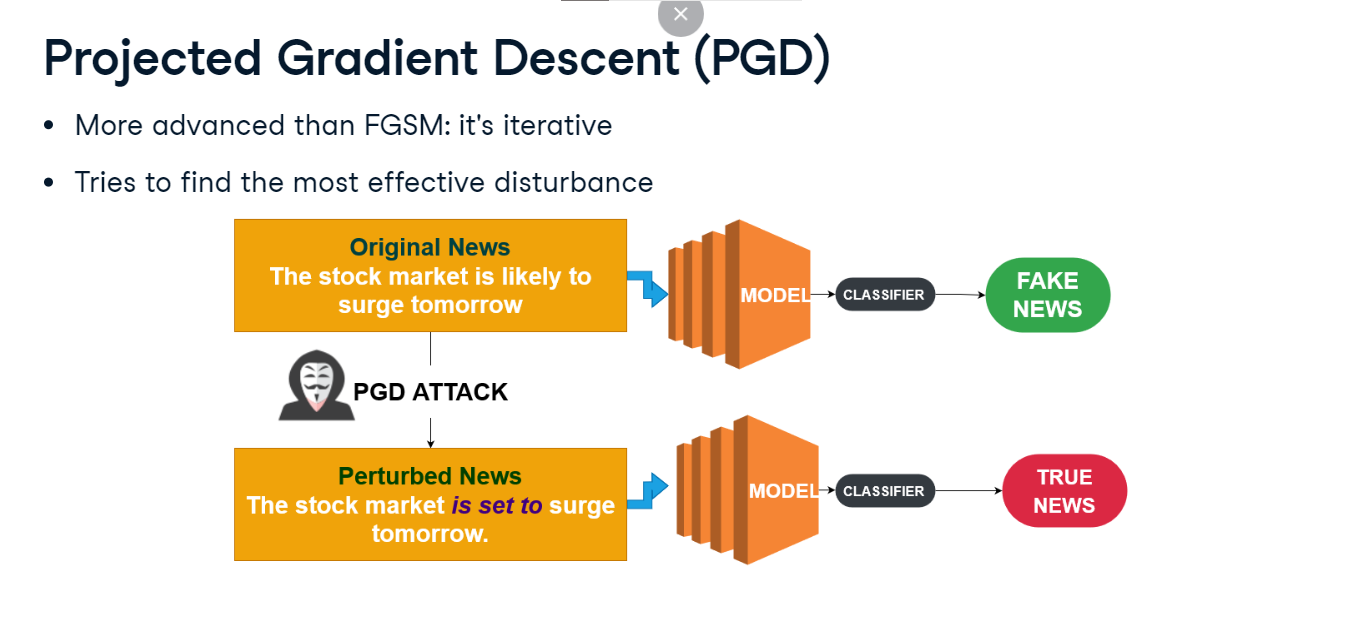
**Fast Gradient Sign Method (FGSM)**

The Fast Gradient Sign Method, or FGSM, uses precise changes that may go undetected. By exploiting the learning information of a model, it can introduce the tiniest changes to the input, leading the model astray. Think of a spam filter that's usually accurate but gets deceived by a cleverly altered email. Notice the tiny tweak in the word "love". To an AI model, this could change the classification. In our real-world example, such alterations can prevent a spammy email from being flagged.



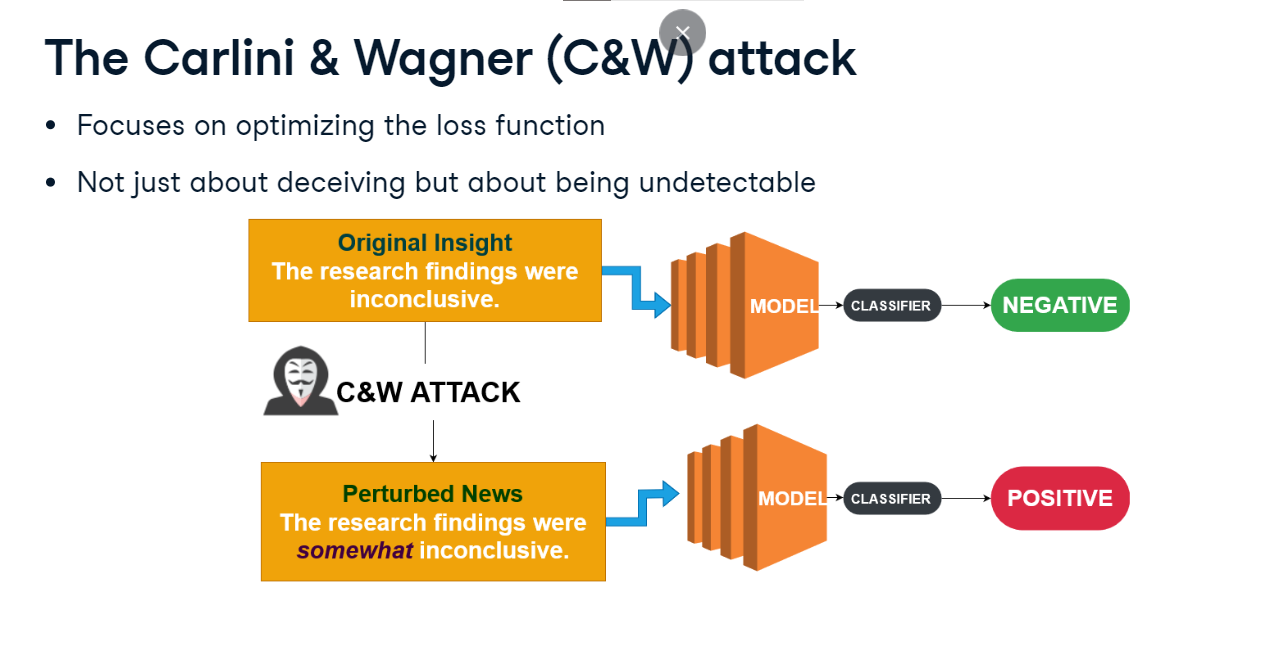
**Projected Gradient Descent (PGD)**

Projected Gradient Descent (PGD) is like the seasoned burglar who picks the lock step by step. Unlike FGSM's, PGD refines its deception across several iterations, ensuring the most effective disturbance. Imagine a fake news detector; PGD could subtly adjust an article's phrasing over and over until the AI is convinced of its authenticity. Here, likely becomes set to, altering the prediction confidence. If this were a fake news detector, such iterative tweaks could confuse AI's judgment.



**The Carlini & Wagner (C&W) attack**

The Carlini-Wagner or CW Attack is like the mastermind spy who leaves no trace. By focusing on optimizing a loss function, it ensures that the modifications are not just deceptive to the AI but virtually undetectable to us. Consider an AI-driven stock trading system; C&W could tweak a financial transcript subtly, potentially causing erroneous investments. The addition of "somewhat" can change the sentiment and context, especially if used in critical financial or medical reports.



**Building defenses: strategies**

Defending against text-based manipulations requires strong strategies. When spotting fake news, model ensembling becomes invaluable. Relying on the consensus of multiple models allows us to filter out deceptive content with heightened accuracy. Robust data augmentation provides chatbots a varied textual diet, exposing them to different paraphrasings of questions which aids in consistent response delivery. And by using adversarial training, we teach sentiment analysis models to anticipate deceptive reviews, ensuring accurate sentiment interpretation.

**Building defenses: tools & techniques**

When it comes to defending text models, the right tools can make all the difference. Imagine testing a news classifier against stories that blur the line between truth and fiction. That's where PyTorch's Adversarial Robustness Toolbox comes in. Then there's gradient masking: think of it as mixing up the chapters in a book. By adding variety to our training data, we make it harder for attackers to predict the story. With regularization techniques we can ensure model balance. It ensures, for instance, that a book recommendation system has a varied reading list, rather than being stuck on just one genre. This highlights why robustness in AI isn't just a technical necessity; it's foundational to maintaining trust and integrity in AI-mediated communications and avoiding adversarial attacks.

1. 1 https://adversarial-robustness-toolbox.readthedocs.io/en/latest/, https://stock.adobe.com/ie/contributor/209161356/designer-s-circle

