# Project Veritas: An AI-Powered Fake Review Detection System for E-Commerce

## Executive Summary

Deceptive product reviews (astroturfing, paid reviews, competitor sabotage) threaten consumer trust and distort market data. Project Veritas proposes a hybrid AI/ML solution combining Natural Language Processing (NLP) with behavioral analysis to automatically detect fraudulent reviews. The system will reduce fake reviews by over 90% within the first year, providing accurate insights to sellers and ensuring customers base their decisions on genuine feedback.

## Problem Statement & Objectives

Challenges include:  
- Deceptive Opinion Spam: False or misleading opinions.  
- Review Spam: Non-opinion based, off-topic, or promotional content.  
  
Objectives:  
- Achieve >95% precision and >90% recall in identifying fake reviews.  
- Reduce fake reviews visible to customers by 80% within 6 months.  
- Provide real-time API classification with <200ms latency.  
- Deliver explainable AI (XAI) outputs for moderator decisions.

## Scope & Features

In-Scope:  
- Review Text Analysis (NLP).  
- Reviewer Metadata Analysis (behavioral).  
- Confidence scoring (0-100%).  
- Moderator dashboard with flagged reviews.  
- Real-time API for review submission checks.  
- Batch processing for periodic re-evaluation.  
  
Out-of-Scope:  
- Automatic review deletion.  
- Seller dashboard integration (phase 2).  
- Image/video content analysis.

## Technical Architecture & Methodology

Data Sources:  
- Public datasets (Yelp, Amazon Product Data).  
- Synthetic adversarial data.  
  
Feature Engineering:  
A. Linguistic Features:  
- Syntax & Grammar (superlatives, pronouns, poor grammar).  
- Sentiment Extremity.  
- Topic Deviation.  
- Embeddings from BERT/RoBERTa.  
  
B. Behavioral Features:  
- Reviewer History.  
- Temporal Patterns.  
- Rating Deviation.  
- Network Analysis.  
  
Model Selection:  
- First Layer: XGBoost for behavioral features.  
- Second Layer: BERT-based Transformer for text.  
- Final Layer: Logistic Regression meta-classifier.  
  
Infrastructure:  
- AWS SageMaker for training.  
- AWS Lambda/API Gateway for real-time API.  
- Amazon S3 for storage.  
- Kubernetes cluster for batch jobs.

## Implementation Phases

Phase 1 (Months 1-3): Data Acquisition, Cleaning, EDA, Baseline model.  
Phase 2 (Months 4-6): Feature Engineering, advanced models (BERT, XGBoost).  
Phase 3 (Months 7-8): Dashboard + Real-time API, A/B testing.  
Phase 4 (Months 9-10): Full deployment, monitoring, feedback loop retraining.

## Evaluation Metrics

Primary Metrics:  
- Precision, Recall, F1-Score, AUC-ROC.  
Business Metrics:  
- Reduction in fake review reports.  
- Improved customer trust scores.  
- Increased conversion rates.

## Ethical Considerations & Bias Mitigation

- Bias Auditing with AI Fairness 360.  
- Adversarial Debiasing during training.  
- Human-in-the-Loop moderation before final action.

## Appendices

- Mock-up of moderator dashboard (to be designed).  
- Sample architecture diagram.  
- Key References:  
 \* Jindal & Liu (2008), Opinion Spam and Analysis.  
 \* Devlin et al. (2018), BERT: Pre-training of Deep Bidirectional Transformers.