Improving LLMs' Classification Ability in Identifying Deceptive Reviews

A small sentence which explains all about this presentation

Created using presentations.ai

Technical Problem

Click here to edit subtitle

01

What are Deceptive Reviews

Deceptive reviews are intentionally misleading and try to sway a product's rating, whether for better or worse.

02

Project Goal

Create a simple yet effective method to classify reviews as truthful or deceptive with LLMs.

03

Utilizing LLMs

The project involves tuning currently existing LLMs to improve classification ability.

Created using presentations.ai

Technical Challenges

Click here to edit subtitle

Deceptive Review Identification

01

Deceptive reviews often resemble truthful ones and happen in many ways (e.g. exaggeration, underplaying flaws, and misleading information). Manual checks for deception are both tedious and time-consuming.

Traditional Detection Methods

02

Detection typically involves sentiment analysis and feature-based approaches with NLP, which can be costly and time-consuming to train, and lack unlabeled deceptive review datasets.

Limitations of LLMs

03

LLMs can be simply accessed from the internet. However they are trained for general purposes and have not been fine-tuned specifically for deception detection, making them less effective in this context.

Created using presentations.ai

Impact and Contributions

01	**Scientific benefits**	Exploring how LLMs handle the detection of deceptive opinions and their reasoning can offer valuable insights into how well they manage subtle language shifts. It also highlights their potential in sentiment analysis.
02	**Availability for small companies**	This approach could provide small companies with a simple and effective way to integrate deceptive review detection into their systems.
03	**Consumer protection**	Deceptive reviews can mislead consumers into buying low-quality products or services. By reducing these reviews, we can protect customers' rights and minimize fraud in online transactions.
04	**Transparency in the marketplace**	Since LLMs are publicly available third-party tools that anyone can use, they could help prevent individuals or companies from manipulating product reviews for personal gain, leading to a more transparent marketplace.

Created using presentations.ai

Limitations of LLMs

01

Struggles with Unfamiliar Patterns

LLMs may struggle when encountering language patterns they weren't exposed to during training, which affecting accuracy in small domains.

02

Performance Instability

LLM performance can be unstable; if the model gets updated, it may yield different results than before.

03

Cost Implications of API Usage

Repeated API calls can become expensive over time which limits both the number of tests and the range of LLMs to use.

Created using presentations.ai

MAiDE-up: Multilingual Deception Detection of GPT-generated Hotel Reviews

- Did a linguistic analysis of real and fake reviews where the fake reviews are generated by the chatGPT.
- Analyse factors including sentiment, location, and language.
- It provides insight into how LLM mimics human language.

Created using presentations.ai

Bad Actor, Good Advisor: Exploring the Role of Large Language Models in Fake News Detection

01

ChatGPT-3.5's Role in Fake News Detection

This paper used ChatGPT-3.5 to examine fake news and found that it struggled with detection accuracy.

02

Adaptive Rationale Guidance Proposal

The author proposes Adaptive Rationale Guidance, which LLM will offer rationales to enhance small models.

Created using presentations.ai

Detecting AI-enhanced Opinion Spambots: A study on LLM-generated Hotel Reviews

01

Usage of LLMs-generated fake hotel reviews

The author utilized LLMs to generate fake hotel reviews for the study.

02

Ease of detection with sufficient training data

It was found that detecting these reviews is relatively easy if there is enough training data available.

03

Distinct vocabulary and style

The vocabulary and style of LLM-generated reviews are notably different from those of human reviews.

Created using presentations.ai

Dataset

400 rows for each category

Balanced dataset across categories

The dataset includes 400 rows each for four categories: truthful positive, truthful negative, deceptive positive, and deceptive negative reviews.

7 platforms

Diverse data sources

Data is collected from seven platforms including Yelp, TripAdvisor, Expedia, Hotels.com, etc.

20 reviews per hotel

Focused on popular hotels

The dataset covers 20 reviews each for the 20 most popular hotels in Chicago in each category.

Real-world reflection

Mirrors competitive market scenarios

Since these hotels are well-known in a big city, the dataset reflects real-world scenarios where fake reviews are more common, especially in highly competitive markets.

Created using presentations.ai

Tools and Systems

Selection of Leading LLMs

01

LLMs like ChatGPT and Gemini are chosen due to their extensive training on massive datasets, making them highly effective for the task.

Access to Comprehensive API Endpoints

02

These LLMs offer comprehensive API endpoints, facilitating easier integration and utilization in various applications.

Powerful Performance

03

ChatGPT and Gemini are currently recognized as the most powerful LLMs, ensuring high performance and accuracy in outputs.

Created using presentations.ai

Experimental Setup

01

Consistent Input Format

- Prepare a prompt that does not change and is used throughout the entire experiment.
- Prepare each row as input in a clear, consistent format.
- E.g. messages: [{ hotel: 'xxx', deceptive: 'true/false', review: 'xxxxx' }]

02

API parameter setting

- Temperature: A higher value will make the response more random and a lower value will make the response more focused and deterministic.
- Max completion tokens: Set this with a proper value to make sure each response fits within a limit to ensure consistency.

03

Structured Output

Provide a JSON output template to the LLM to ensure consistent output format and reliable type safety.

Created using presentations.ai

Evaluation Metrics

Evaluating the Effectiveness of LLMs in Detecting Fake Reviews

Accuracy

Performance Overview

Evaluated by the percentage of correct classifications out of the total number of samples.

Precision of deceptive reviews

Minimizing False Positives

Shows how many reviews classified as deceptive are actually deceptive.

Recall of deceptive reviews

Minimizing False Negatives

Tells us the percentage of the actual deceptive cases that were successfully identified by the model.

Created using presentations.ai

Early Stage Experiments

Zero-shot Testing

GPT-4o correctly identified 4 out of 20 deceptive reviews about the same hotel.

01

02

Ten-shot Testing with Promt Engineering

- Given 5 truthful and 5 deceptive examples in advance.
- Tested on a different hotel with 10 truthful and 10 deceptive reviews.
- Correctly identified 12/20 reviews.

Next Step

Fully evaluate the performance of GPT-4o and start fine-tuning.

03

Created using presentations.ai

Thank You!

A small sentence which explains all about this presentation

Created using presentations.ai