Spam/Fake Review Detection

November 19, 2024

Yusheng Yang, Shelton Zhou, Chuyang Wang

1 Abstract

Fake online reviews undermine trust, mislead consumers, and create unfair competition. This proposal focuses on developing a machine learning system to detect and prevent fraudulent reviews, helping platforms ensure authenticity and fairness. Using advanced algorithms, the model will adapt to evolving tactics, providing reliable, long-term results. By restoring trust in online feedback, this project supports ethical business practices, empowers consumers to make better decisions, and fosters a transparent digital marketplace.

2 Background/Motivation

Problem is defined as **detecting fake or spam reviews to address the growing issue of misleading users in making decisions based on fabricated feedback**. This is particularly important in domains like hotel bookings or online marketplaces where reviews significantly influence customer trust.

This is a problem that can be understood by anyone due to its practical implications as 95% of consumers read reviews before purchases, and 58% are willing to pay more for well-reviewed products (Global Newswire). Fraudsters use fake reviews to manipulate perceptions, promote products, or harm competitors, leading to financial losses, dissatisfaction, and eroded trust in platforms. By focusing on using deep learning, this project aims to automate and improve the accuracy of fake review detection compared to traditional rule-based methods.

While some studies had similar interests, the problem is not defined exactly the same. Traditional methods like rule-based systems and classical machine learning (e.g., Naïve Bayes, SVM) struggle to capture semantic and contextual nuances. While deep learning models like CNNs and LSTMs have shown promise, challenges like accuracy and scalability persist. Ott [1] introduced a dataset of fake and real hotel reviews, highlighting the challenges of detection. This project addresses these gaps by combining multiple deep learning methods with two diverse datasets to ensure robust evaluation and scalability.

Solving this problem is critical to protecting users, building trust in platforms, and ensuring fair competition. Automated systems powered by deep learning are essential for handling the growing volume of reviews.

The proposed approach integrates CNN for feature extraction, LSTM for sequence understanding, and Logistic Regression as a final classifier to optimize accuracy. By leveraging two datasets, this project should offer a comprehensive analysis and concludes with a comparison of model effectiveness.

3 Proposed Implementation

3.1 Datasets

Two publicly available datasets will be used for this project:

1. **20 Chicago hotels dataset** [1],[2]
   https://www.kaggle.com/datasets/rtatman/deceptive-opinion-spam-corpus
   It includes 400 truthful positive reviews, 400 deceptive positive reviews, 400 truthful negative reviews and 400 deceptive negative reviews from Expedia, TripAdvisor, Yelp and so on.

2. **Fake Reviews Dataset** [3]
   https://www.kaggle.com/datasets/mexwell/fake-reviews-dataset
   It contains 20,000 authentic reviews and 20,000 fake reviews generated by computer.

3.2 Design Requirements

The project will develop the 3 following model architectures:

- Bidirectional LSTM with GLoVE 50D embeddings.
- CNN-LSTM with Doc2Vec and TF-IDF.
- Logistic Regression with TF-IDF.

Test accuracy will be computed for evaluation purposes.

3.3 Plans for architectures and designs

The following models will be implemented:

1. **Bidirectional LSTM with GLoVE 50D Word Embeddings**: leverages semantic relationships in text for sequential modeling.
2. **CNN-LSTM with Doc2Vec and TF-IDF**: combines convolutional layers for feature extraction and LSTM layers for sequential understanding, with TF-IDF for weighted term importance.
3. **Logistic Regression with TF-IDF**: A lightweight model serving as a baseline for comparison.

3.4 Step-by-step process and success metrics

1. **Data preprocessing**: clean, tokenize, generate word embeddings and TF-IDF features, create training and test splits.
2. **Implementing models**
3. **Evaluation and comparison**: measure accuracy, identify the best-performing model based on metrics and computational efficiency.

3.5 Task Breakdown

- **Yusheng Yang:**
  - Responsible for dataset preprocessing and feature engineering.
  - Implement TF-IDF vectorization for Logistic Regression and CNN-LSTM models.
- **Shelton Zhou:**
  - Develop and train the Bidirectional LSTM model using GLoVE embeddings.
  - Tune hyperparameters for optimal performance.
- **Chuyang Wang:**
  - Implement CNN-LSTM with Doc2Vec and TF-IDF.
  - Handle result analysis and model comparison using accuracy and other metrics.

4 Feasibility and Limitations

4.1 Endpoint Goals

The project aims to achieve an accuracy of at least 95%, which is considered sufficient for its objectives. Beyond this threshold, improving accuracy further would demand significantly more computational resources. Given the trade-off between resource investment and potential gains, pursuing higher accuracy may not be worthwhile.

4.2 Challenges and Initial Solutions

The biggest challenge is the limited availability of computational resources. To address this, the project proposes leveraging educational credits from AWS or Google Colab.

4.3 Limitations

A potential limitation is the risk of failing to achieve the 95% accuracy target due to factors that are not yet fully understood. This includes uncertainties around dataset quality, model architecture, or problem complexity. The project will explore these areas further to identify potential constraints and determine whether they impose fundamental limits on performance.

5 Potential Impact

Detecting spam and fake online reviews can significantly enhance trust and credibility in online platforms by ensuring that reviews reflect genuine customer experiences. This not only boosts consumer confidence but also strengthens the platform's reputation for fairness and transparency. For consumers, it enables more informed purchasing decisions, free from the influence of manipulated ratings, while also protecting them from scams or low-quality products promoted through deceptive reviews. Businesses, especially smaller ones, benefit from a level playing field where genuine feedback allows them to compete based on the quality of their offerings, while unethical practices like inflating reputations or harming competitors through fake reviews are discouraged.

6 References

[1] M. Ott, Y. Choi, C. Cardie, and J.T. Hancock. 2011. Finding Deceptive Opinion Spam by Any Stretch of the Imagination. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies.

[2] M. Ott, C. Cardie, and J.T. Hancock. 2013. Negative Deceptive Opinion Spam. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

[3] Salminen, J., Kandpal, C., Kamel, A. M., Jung, S., Jansen, B. J. (2022). Creating and detecting fake reviews of online products. Journal of Retailing and Consumer Services, 64, 102771. https://doi.org/10.1016/j.jretconser.2021.102771