CSC 522 ALDA Project

Online Review Spam Detection

Presented by:

Harsh Kachhadia (hmkachha)

Tirth Patel (tdpatel2)

Piyush Biraje (pbiraje)

Introduction

- Before buying any product or service online, it is very common for us to check its reviews
- In this project, we have used machine learning models to predict if a given review is genuine or not
- Focused on hotel reviews published on websites such as TripAdvisor, Yelp, Expedia, etc.

★☆☆☆☆ **Built well but disconnects regularly**
Reviewed in the United Kingdom on October 14, 2019
Color: Red switch | Verified Purchase
Thinking of returning this item not sure if it's faulty built well but every 30 seconds its disconnecting off bluetooth not tired hardwired but I only bought it for the wireless side I already have a hard wired mechanical keyboard

Craig Spencer
★★★★★ **I am sleep-deprived from not being able to put down this book!**
Reviewed in the United States on September 23, 2020
Verified Purchase
Deception Point is an apt name. The mind blowing discovery, political schemes that read like present day, heroes, villains, special forces fast action, secret tech, all make a great plot. Dan Brown never dissapoints and makes me want to research the actual technology details weaved in the story. Will read again!

Can you tell if these reviews are real or fake?

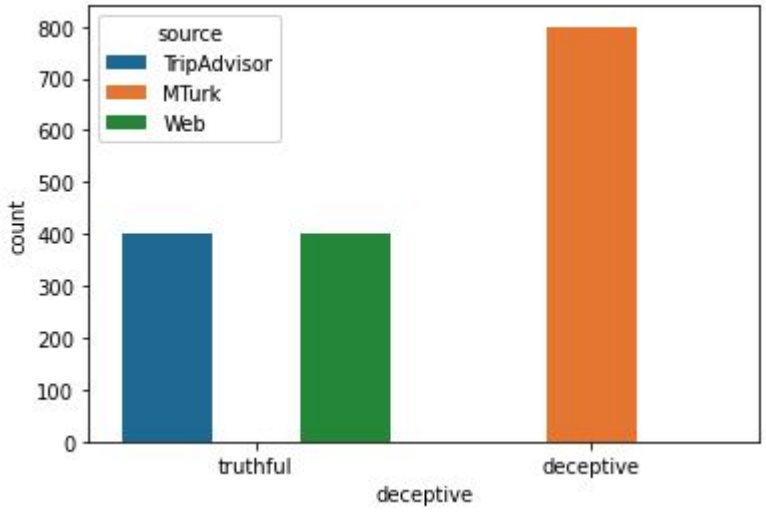
Method: Data

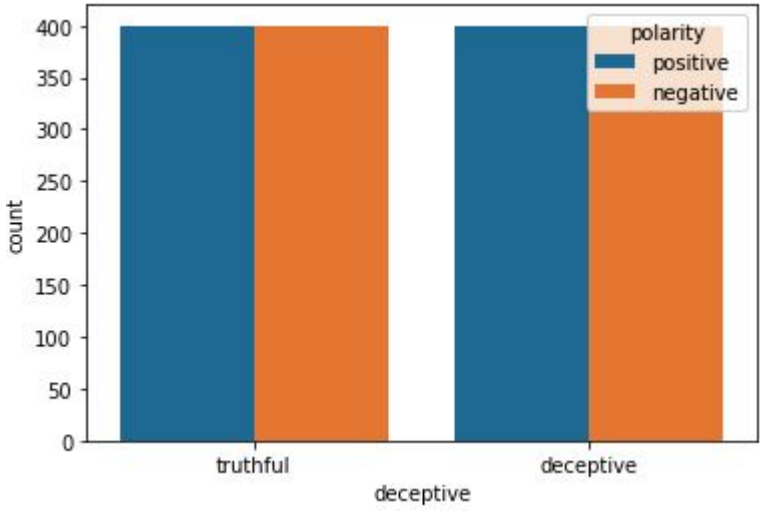
- **Phase 1: Hotel review dataset obtained online**

  Source: Kaggle, Hotel reviews from 20 chicago hotels

  Attributes: Class Label (Deceptive/Non-Deceptive), Hotel Name, Polarity of review, Source/Website, Reviews (text)

- **Some info on dataset:**
  - 400 Truthful positive polarity reviews from TripAdvisor.
  - 400 Truthful negative polarity reviews from other websites like expedia, yelp, etc.
  - 800 Deceptive (both positive and negative polarity) reviews from Mechanical Turk.





Method: Data Preprocessing

- **Data Preprocessing:** Process the reviews using **NLTK library.**

- **Steps involved in data preprocessing are as follows**:
  1. Convert text to lowercase
  2. Remove - punctuation, non-alphanumeric characters, hyperlinks, special characters, new line characters, digits, extra space
  3. Removal of stopwords.

Sample text with Stop Words	Without Stop Words
GeeksforGeeks – A Computer Science Portal for Geeks	GeeksforGeeks , Computer Science, Portal ,Geeks
Can listening be exhausting?	Listening, Exhausting
I like reading, so I read	Like, Reading, read

Methodology

Goal: To obtain a labelled hotel reviews dataset and identify the best classifier

Phase 1:

1. Get the Kaggle data corpus & pre-process it
2. Train Classifiers on this data
3. Select the best classifier

Phase 2:

1. Gather new data using web scraper
2. Preprocess this data
3. Label the data using the selected classifier
4. Add polarity to the dataset to make it more useful

Experimental Setup: Classifier Models

- Converted the pre-processed text in the "Text" column to TF-IDF feature vectors, then applied the following models:
  - **Multinomial Naïve Bayes:**

    Sklearn - MultinomialNB Classifier (default hyperparameters) on the TF-IDF feature vectors.
  - **Logistic Regression:**

    Sklearn - LogisticRegression(n_jobs = 1, C = 1e5)
  - **SVM (Linear SVC):**

    LinearSVC(loss='hinge', C=5, random_state=42)
  - **Stochastic Gradient Descent:**

    SGDClassifier (loss = 'hinge', penalty = 'l2', alpha = 1e-3, random_state = 42, max_iter = 5, tol = None)
  - **Random Forest:**

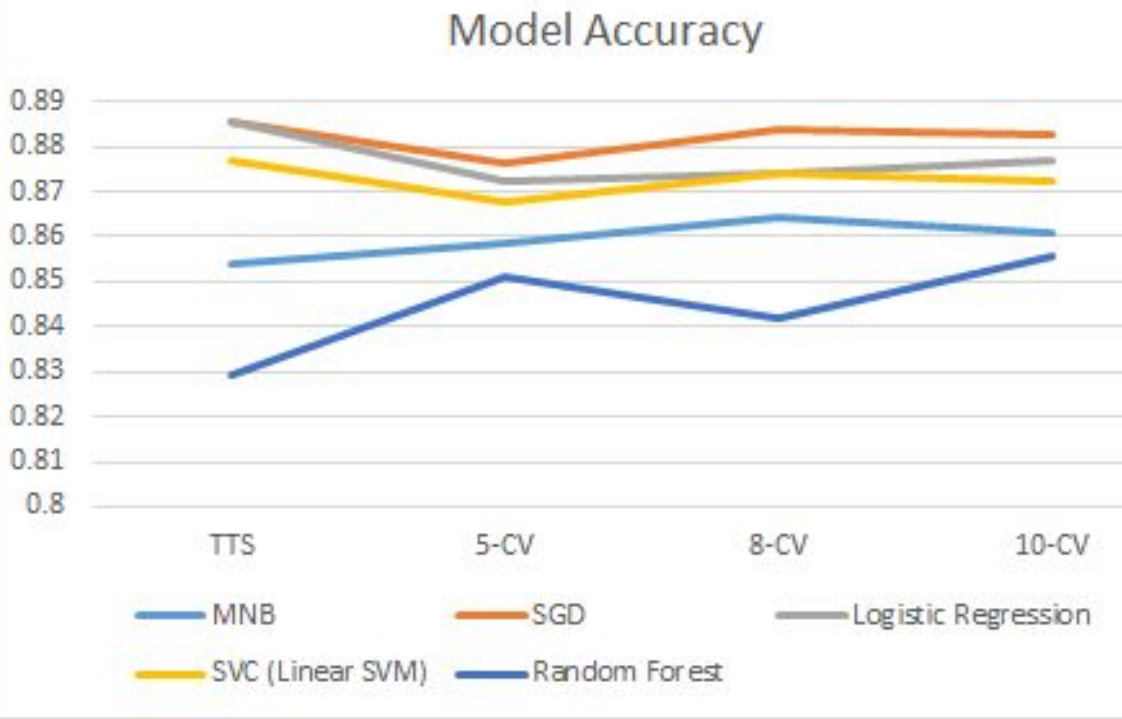
    Random Forest Classifier with default hyperparameters

Experimental Setup: Model Evaluation

- **Model Evaluation Methods implemented:**
  - **K-Fold CV** : 5-CV, 8-CV, 10-CV
  - **Holdout Method** (train, test - split) (0.7-train, 0.3 test)

- **Evaluation metrics:**
  - Accuracy
  - Precision
  - Recall
  - F-measure

MidTerm Results



- Stochastic Gradient descent (SGD) algorithm has returned the best results with consistently highest accuracy (88.5%) across all model evaluation techniques. (**Why Accuracy??**)
- For our application, the SGD method of classification is best suited

Experimental Setup: New Dataset Generation

Targeted City : Las Vegas

Hotels:

We targeted top 20 hotels of Las Vegas listed as per USNews 2020: Top 20 Hotels in Las Vegas

Why top 20?

As per statistics, most popular hotels are the ones most targeted by such Spam/deceptive reviews.

Experimental Setup: New Dataset Generation

Sources from which Hotel Reviews were scraped:

- Trip Advisor
- Expedia
- Yelp

Reason for selecting them:

- **TripAdvisor** and **Expedia** are some of the most popular & trusted sites for hotel booking → that is where people will look for reviews.

- **Yelp** - It is the most common review site where people write reviews about various places including hotels (for diversification of data)

Experimental Setup: New Dataset Generation

Web Scraping:

3 team members -> each built 1 -> **3** web scrapers in python (TripAdvisor, Expedia, Yelp)

Python Libraries utilized:

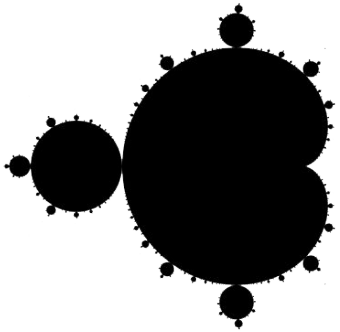
- **BeautifulSoup**
- **urllib**
- **Pandas**
- **Sklearn**

Experimental Setup: New Dataset Generation

Data Tagging

Web scraped reviews were **preprocessed** & **labelled** with **SGD Model (**selected for its best performance**)** trained on **Kaggle Dataset** (mentioned in previous slides)

Polarity of reviews was labelled with **TextBlob Sentiment Analyzer** in Python.



TextBlob

Results: Final Dataset Statistics

- **Dataset:** (Link to our Dataset)
  - **Sources:** Expedia, TripAdvisor, Yelp
  - **Records:** 1200 (20 Hotels * 20 Reviews * 3 Sources)
  - **Label Counts:** 815 **Truthful** , 316 **Deceptive**
  - **Review Polarity Counts** :
    Negative: 1072, **Positive:** 116, **Neutral:** 12

Conclusion and Take Home Message

- One of the major problems that we faced in this project is the lack of large enough data for spam detection
- We observed that there are very few openly available corpus for hotel review that are labelled with truthful / deceptive.
- In this project, we have created deceptive opinion spam dataset, which will be helpful for future projects, as this will be openly available.
- Also, with our SGD Classifier, we can now successfully test the genuinity of an online hotel review with 88.5% accuracy.

Important Links & References

Important Links:

Kaggle Chicago Hotel Review Dataset

References

- https://www.aclweb.org/anthology/P11-1032.pdf
- https://www.aclweb.org/anthology/N13-1053.pdf
- https://www.cs.uic.edu/~liub/publications/reviewSpam-2007.pdf
- https://arxiv.org/pdf/1903.12452.pdf