BT4222 PROJECT PROPOSAL

Group 15

Jasvinder Singh	A0155000A
Shen Yuting	A0213633N
Xiao Zuoling	A0162759R
Zhou Bangda	A0171806E

1. Problem Description

This project focuses on building an automated system to detect fake restaurant reviews on Yelp.

In this day and age of vast internet connectivity, more people place a high importance on online reviews before patronizing a business. 90% of consumers check the reviews of a business before visiting it and 88% believe the reviews provided online (Saleh, 2019). From a business standpoint, these statistics may create motivation for businesses to provide fake reviews to sway potential customers to patronize the business. As such, it is important that the credibility of reviews on such sites can be validated so as to give consumers an impartial and reliable impression of businesses that they wish to explore. It is practically difficult for humans to manually filter out the fake reviews. The low speed and accuracy of detection by humans makes automation important. We will be using natural language processing and machine learning techniques to tackle this problem.

2. About the Data

For this project we are using Yelp's restaurant review database. The database contains 3 tables: review, reviewer and restaurant. For this fake review detection project, we are using the review and the reviewer datasets. The dataset was crawled from Yelp by previous researchers. Our data source comes from a professor at the University of Illinois in Chicago, who provided us with a password to access the database.

The review data source contains columns:

Column name	Date Type	Description
date	object	The date when the review was posted
reviewID	object	The unique review ID
reviewerID	object	The unique reviewer ID
reviewContent	object	The content of review
rating	int64	Review rating from 1 to 5
usefulCount	int64	The number of 'useful' votes received
coolCount	int64	The number of 'cool' votes received
funnyCount	int64	The number of 'funnyl' votes received
flagged	object	NR: fake review; YR: real review
restaurantID	object	The unique ID of restaurant being reviewed

There are 721,452 unique reviews in the table written by 15,941 unique reviewers. The reviews are collected from October 2004 to October 2012. Of all the reviews, 402,t774 (55.8%) are labeled as fake reviews and 318,678 (44.2%) are labeled as real reviews as indicated by the flagged column. Overall, the review classes are quite balanced.

The flagged column is obtained by Yelp's review filtering engine. Every Yelp review is evaluated by Yelp’s recommendation software based on quality, reliability and user activity. In order to prevent both negative and positive review fraud, Yelp’s software takes into account user’s profile, engagement level, IP address, length of the review, uniqueness of the content and business factors.

The reviewer table contains:

Column Name	Type	Description
reviewerID	object	The unique ID of reviewer
name	object	The encode name of reviewer
location	object	The location of reviewer
yelpJoinDate	object	The date when the user registered on Yelp
friendCount	int64	The No. of friends user has
reviewCount	int64	The No. of reviews user has posted
firstCount	int64	The No. of times the user’s review ranked first
usefulCount	int64	The No. of useful votes the user received
coolCount	int64	The No. of cool votes the user received
funnyCount	int64	The No. of funny votes the user received
complimentCount	int64	The No. of compliments votes the user received
tipCount	int64	The No. of tips the user give
fanCount	int64	The No. of fans the user has

There are 16,941 unique reviewers in the dataset. 9.65% of them only posted 1 review on Yelp. About 49% of the reviewers posted more than 10 posts. And 10% of them posted more than 100 reviews.

The reviewer table may be interesting to analyze because the fake reviewers tend to share some common characteristics such as giving large amounts of 5 stars reviews with no negative comments (Hill, 2018).

3. Exploration of Features

From the various columns present in the dataset, there are some basic features that would be interesting to explore.

Feature	**Explanation**
Number of Friends/Reviews	The number of friends a reviewer has may suggest whether the reviewer is fake or not because fake reviewers are not likely to be active in the Yelp community
Text Length	Length of the review text may have a correlation with the legitimacy of the review as fake reviewers may post shorter reviews.
“Useful”, “Funny”, “Cool” Remarks	Feedback from other users who have referred to the review would be important in determining the reliability of a specific review.
Time Between Reviews	Fake reviewers may post reviews in short intervals of time because of the potential sporadic opportunities to earn money. Thus, the time between reviews may be shorter for fake reviewers.

4. Model Objective and Hypotheses

From the dataset, we would like to train a model which is able to make decisions on whether a review is fake based on the review itself and the reviewer’s profile. We also want to figure out which variables or features are important to determine if a review is fake.

In addition, we want to test some of the hypotheses related to fake reviews and the reviewers. Our hypotheses are that fake reviews belonging to the same reviewer will (i) have higher similarity and (ii) fake reviews tend to exhibit strong emotions.

5. Textual Machine Learning Methodology

Part I: Text Analysis

We will perform feature engineering the text review for our prediction. Hence, we list down some potential textual analyses and its corresponding methods:

A. *Pre-processing*

Before feeding the text body (review content) into any machine learning model, it needs to be pre-processed into word vectors. For this project, we are going to follow a standard text pre-processing pipeline using SpaCy (Balatsko, 2019).

Firstly, we split the text into tokens via tokenization. Secondly, we clean the text tokens to remove unwanted content. This will involve punctuation removal and stop words removal. Next, normalize the data by converting dates, numbers and signs to text, converting any non-text information into textual. Lastly, lemmatize the tokens to become the root words to reduce the inflectional forms.

B. *Embedding*

We are going to try the following embeddings: Bag of words, TF-IDF, Word2Vec (developed by Google), Glove (developed by Stanford)

C. *Similarity between different reviews from the same user*

As posting similar review might be a strong indicator for paid-review user, we are going to compare the similarity across the reviews being posted by the same user. we will use the following methods suggested by a few papers and articles online:

a. Jaccard Similarity
b. Different embeddings + K-means
c. Different embeddings + Cosine Similarity
d. Different embeddings + Latent semantic analysis + Cosine Similarity

D. *Sentiment of a certain text*

Fake reviews are usually linked to strong emotions and come out with extreme ratings like 1 or 5 stars (Beaton, 2018).

So sentiment analysis can serve as an effective tool to provide the polarity to the text and classify it into positive and negative with different levels. In addition, aspect mining like part-of-speech tagging can be used with sentiment analysis to obtain information from the reviews by identifying different aspects of the text (Vaidya, 2018).

Part II: Classification Problem

Essentially, by predicting whether a review in the test dataset is a fake review, we are trying to solve a supervised binary classification problem, hence we will use the following machine learning methods to reach the final conclusion are:

A. *Logistic Regression*

Logistic Regression is a widely used binary classification model. It uses a logistic function to model a binary dependent variable, which is well suited for our goal.

B. *Naive Bayes*

Naive Bayes is a probabilistic classifier based on Bayes' theorem, and it's based upon the naïve assumptions that the features in a dataset are mutually independent.

C. *Support Vector Machine*

Support Vector Machine is a widely used model for classification. SVM tries to make a decision boundary in such a way that the separation between the two classes is as wide as possible.

D. *Random Forest*

Random Forest is an ensemble learning model for classification by developing many decision trees. It uses averaging to improve the predictive accuracy and control over-fitting

E. *XGBoost*

XGB is a popular library providing the gradient boosting models that is widely used in Kaggle competition. It has been very popular in recent years due to its versatility, scalability and efficiency.

References

Balatsko, M. (2019, May 21). *Text preprocessing steps and universal reusable pipeline*. Retrieved from TowardsDataScience: https://towardsdatascience.com/text-preprocessing-steps-and-universal-pipeline-94233cb6725a

Beaton, C. (2018, June 13). *Why You Can't Really Trust Negative Online Reviews*. Retrieved from The New York Times: https://www.nytimes.com/2018/06/13/smarter-living/trust-negative-product-reviews.html

Hill, C. (2018, December 10). *10 secrets to uncovering which online reviews are fake*. Retrieved from MarketWatch: https://www.marketwatch.com/story/10-secrets-to-uncovering-which-online-reviews-are-fake-2018-09-21

Saleh, K. (2019). *The Importance Of Online Customer Reviews*. Retrieved from Invesp: https://www.invespcro.com/blog/the-importance-of-online-customer-reviews-infographic/

Vaidya, N. (2018). *5 Natural Language Processing Techniques for Extracting Information*. Retrieved from AureusAnalytics: https://blog.aureusanalytics.com/blog/5-natural-language-processing-techniques-for-extracting-information