Supplementary Materials:

Node Embedding Approach for Accurate Detection of Fake Reviews: A Graph-Based Machine Learning Approach with Explainable AI

Dataset Description

In this study, we employed two publicly available datasets, the Deceptive Opinion Spam Corpus (OpSpam) and YelpChi. Table 1 presents the list of features, their data types, and descriptions. Furthermore, Table 2 displays the sample reviews for each class from the datasets.

Table 1: The description of the features of the Deceptive Opinion Spam Corpus dataset.

Feature	Data Type	Description
hotel	string	The name of the reviewed hotel.
text	string	The review written for the hotel.
polarity	category	The polarity of the review defined either as positive or negative.
source	string	The source of the review.
class	nominal	The class of the review (0—deceptive; 1—truthful)

Table 2: The sample reviews from the OpSpam dataset (0—deceptive; 1— truthful).

Reviews	Class
This hotel is the perfect location for downtown Chicago shopping. The only thing is the pool is extremely small - it is indoors, but looks much larger on the website.	**1**
Well located and well staffed. The Amalfi was a clean and comfortable place to stay. It lacks a restaurant or bar but many are near by. Free continental breakfast and evening cocktail.	**1**
Great hotel with beautiful scenery! The staff was wonderful and the rooms were comfortable and spacious. They even had a docking station for my iPod!	**0**
I stayed at Swissotel Chicago when I was on business and it was very nice. The staff was very helpful and room was very clean. I would stay again in a heart beat!	**0**

YelpChi dataset comprises of 67,395 hotel and restaurant reviews from Chicago, which are divided into separate datasets: YelpChi Hotel and YelpChi Restaurant. The dataset contains reviews from 201 hotels and restaurants by 38,063 reviewers. The datasets are highly imbalanced with the proportion of fake reviews consisting of a little over 13% (see Table 3).

Table 4 presents the sample reviews for each class in the YelpChi dataset.

Table 3: The statistics of the YelpChi dataset.

Name	Total Sample	Fake Reviews		Truthful Reviews	
		Total	%	Total	%
YelpChi Hotel	5,854	778	13.29	5,076	86.71
YelpChi Restaurant	67,395	8,141	13.23	53,400	86.77

Table 4: The sample reviews from the YelpChi dataset (0—deceptive; 1— truthful).

Reviews	Class
Very tired hotel...but great location.	1
Beautiful hotel in a great part of Chicago.	1
Small but nice room. Free internet. Great staff.	0
Great hotel - MUCH better crowd on weeknights	0

Exploratory Data Analysis

This study conducts exploratory data analysis (EDA) to gain a deeper understanding of the data, reveal hidden patterns and insights, and detect possible correlations and trends that could be beneficial in steering additional analysis and modeling for the OpSpam dataset via visualizations.

Figure 1 presents a comparison of the distribution of LIWC (Linguistic Inquiry and Word Count) between truthful and deceptive reviews that have a difference of more than 15%. LIWC can provide insights into the linguistic differences between the two classes. Deceptive reviews contain a significantly higher frequency of first-person pronouns (“i”), cognitive processes (“insight”) that include words such as ‘think’ and ‘know’, and social processes (“family”) with words such as ‘wife’ and ‘husband’. On the other hand, truthful reviews contain a higher frequency of perceptual processes (“hear”), negative affective processes (“sad”), and informal language (“assent”) which include words such as ‘agree’ and ‘heard’.

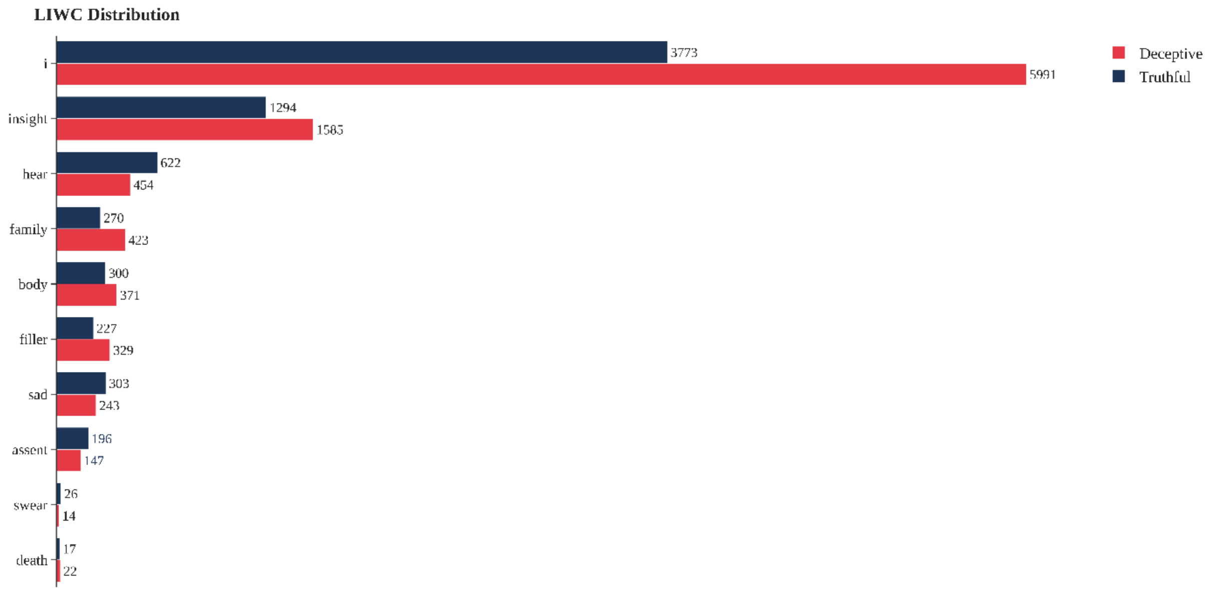

Figure 1: LIWC Distribution between Truthful and Deceptive Reviews.

Figure 2 shows the part-of-speech (POS) tagging distribution between truthful and deceptive reviews. POS tagging helps us understand the difference in language use between the classes. The POS tagging distribution suggests that truthful reviews tend to have a higher frequency of nouns, determiners, full stops, and adjectives, while deceptive reviews tend to have a higher frequency of verbs, adverbs, and pronouns.

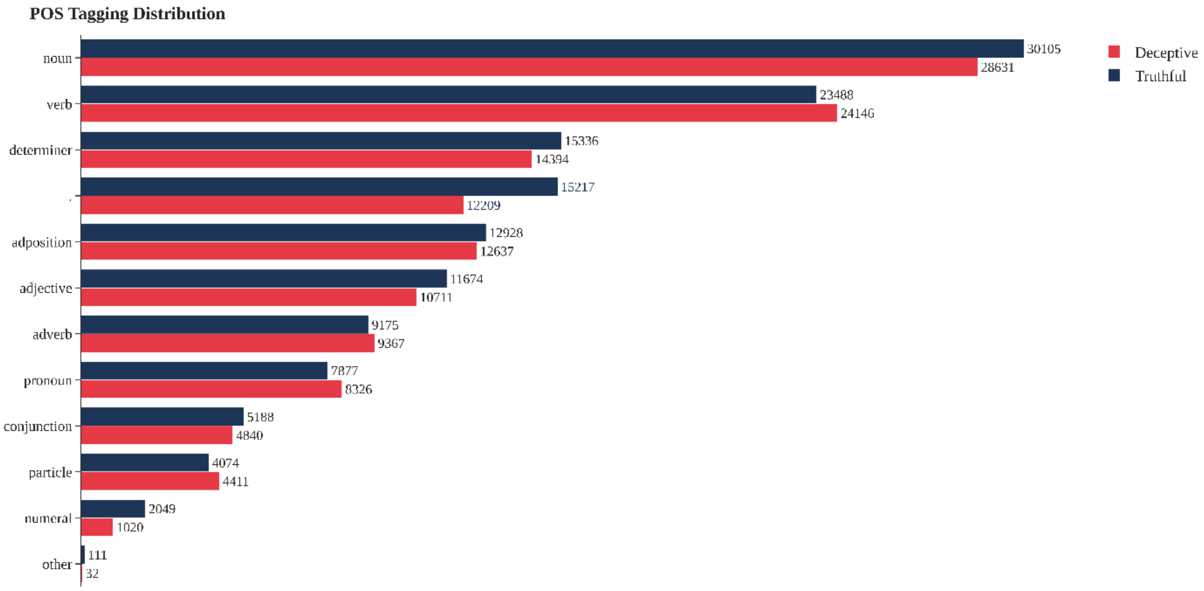

Figure 2: POS Tagging Distribution between Truthful and Deceptive Reviews.

Distance Matrix

In graph theory, a distance matrix is a square matrix containing the distances between the elements of a set. If there are *M* elements, this matrix size will be *M***M*. Here is an example of matrix:

$$A = \begin{bmatrix} a_{11} & a_{12} \dots & a_{1n} \\ a_{21} & a_{22} \dots & a_{2n} \\ a_{n1} & a_{n2} \dots & a_{3n} \end{bmatrix}$$

Where value a_{ij} equals the number of edges from the vertex *i* to *j*. There are lot of metrics techniques to calculate distance between two values. For example, cosine similarity, Euclidean distance, Manhattan distance, Jaccard distance, hamming distance and dot product are used to calculate distance matrix. In this study, we have used nine distance metrics for calculating distance matrix in the best selected features. There are Euclidean distance, cosine similarity, hamming distance, cityblock (Manhattan) distance, Jaccard distance, Minkowski distance, Canberra distance, Chebyshev distance, and braycurtis distance.

Euclidean distance: $d_{ij}^2 = \sum_k \left(x_{ik} - y_{jk}\right)^2$

Where *d* is the distance measure and (*x*,*y*) is the data points.

Cosine similarity: Similarity $= \frac{(A \cdot B)}{(\|A\| \cdot \|B\|)}$

Where *A* and *B* are the two vectors. *A*·*B* is the dot product of two vectors. $\|A\|$ is the L2 norm of *A*.

Manhattan distance: $d = \sum_{i=1}^{m} |x_i - y_i|$

Where *x* and *y* are the data points.

Minkowski distance: $d_{ij}^2 = \sum_{i=1}^{n} |x_i - y_i|$

Where *d* is the distance and (*x*,*y*) are the data points. By using this formula, we find the distance between two points.

Hamming distance: $D = x \oplus y$

Where *D* is the distance and (*x,y*) are the data points.

Chebyshev distance: $d = \max_{i} |x_i - y_i|$

Where *d* is the distance and (*x,y*) is the data points. The equation takes the maximum absolute value of the difference between the points.

Canberra distance: $d = \sum_{i=1}^{n} \frac{|x_i - y_i|}{|x_i| + |y_i|}$

Where *d* is the distance and (*x,y*) is the data points. It is a weighted version of Manhattan distance.

Jaccard similarity: $d = \frac{(A \cap B)}{A \cup B}$

Braycurtis distance: $d = 1 - \frac{2C_{ij}}{S_i + S_j}$

Where *d* is the distance, *c* is the sum of lesser value for the specimens, *s* is the number of specimens.

Adjacency matrix

An adjacency matrix is a representation of the graph where the rows and columns are the nodes of the graph. It is also called as connection matrix. It is represented as follows:

$$Adjacency\ Matrix = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

The diagonals contain zero values. If the graph has no self-loops, the adjacency (vertex) matrix should have zero diagonal values. The main diagonal contains 0s and has no information. The upper triangular part of the matrix is just a mirror of the lower triangular matrix. In this study, we set threshold value to round off the distance matrix using mean of the matrix. Then above the threshold value, we label the matrix values as same matrix value and below the threshold value as ‘0’. Finally, we created a weighted adjacency matrix based on this threshold value.