Results and Evaluation: Humor Identification Model

June 8, 2025

1 Overview

The project uses the Yelp Academic Dataset, from which 10,000 reviews were selected (5,000 humorous and 5,000 non-humorous). After preprocessing and outlier removal, 8,068 reviews were used with the following splits:

- Training set: 4,840 reviews
- Validation set: 1,614 reviews
- Test set: 1,614 reviews

  Novel Application of Zagreb Indices: Utilizes both traditional Zagreb indices and their Upsilon variants to capture the structural properties of text represented as semantic graphs

  Enhanced BERT Integration: Implements a fine-tuned BERT model with balanced class training and gradient accumulation

  Ensemble Learning: Combines multiple machine learning approaches (SVM, Naive Bayes, MLPs, BERT) for optimal performance

2 Methodology

2.1 Text Preprocessing

- Custom tokenization with stopword removal
- Outlier removal using 10th and 90th percentiles of token lengths
- Embedding generation using Word2Vec and GloVe

2.2 Feature Engineering

- **Graph-Based Features**: Conversion of text to semantic graphs with multiple window sizes
- **Zagreb Indices**: (i) 9 Traditional Zagreb indices (First Zagreb, Second Zagreb, co-indices, etc.)

  (ii) 3 Upsilon Zagreb indices capturing deeper structural properties

- **Stylistic Features**: Capitalization, punctuation, word length, humor-related words
- **Embedding Features**: Statistical aggregations of word embeddings

2.3 Model Architecture

The system employs an ensemble of various classifiers:

1. Support Vector Machine (SVM) with RBF kernel
2. Gaussian Naive Bayes
3. Multi-Layer Perceptron with Adam optimizer
4. Multi-Layer Perceptron with RMSprop optimizer
5. Stacking Ensemble (Random Forest, Gradient Boosting, Logistic Regression)
6. Fine-tuned BERT model

3 Model-wise Evaluation Metrics

3.1 Overall Ensemble Performance

Metric	Score
Accuracy	83.83%
F1 Score	83.13%
Precision	87.48%
Recall	79.19%

Table 1: Final Ensemble Classifier Performance on Test Set

3.2 Detailed Model Metrics

Model	F1 Score	Accuracy
Support Vector Machine (SVM)	78.89%	79.80%
Naive Bayes	79.10%	78.75%
MLP Classifier (Adam)	77.77%	78.50%
MLP Classifier (RMSprop)	78.11%	79.06%
Stacking Ensemble	80.60%	81.60%
BERT (Fine-tuned)	80.00%	80.55%
Final Weighted Ensemble	**83.13%**	**83.83%**

Table 2: Model-wise F1 and Accuracy Scores

4 Graphical Performance Comparison

4.1 F1 Scores, Accuracy and Confusion Matrix

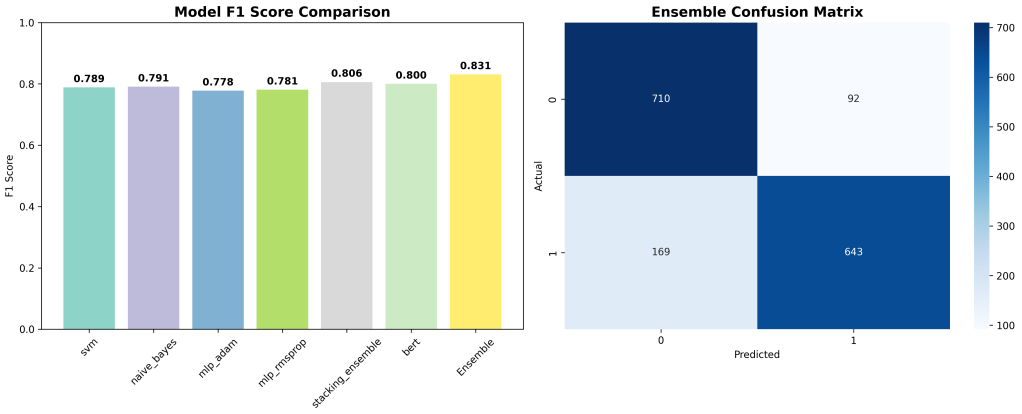

Figure 1: F1 Score and Confusion Matrix Across Models

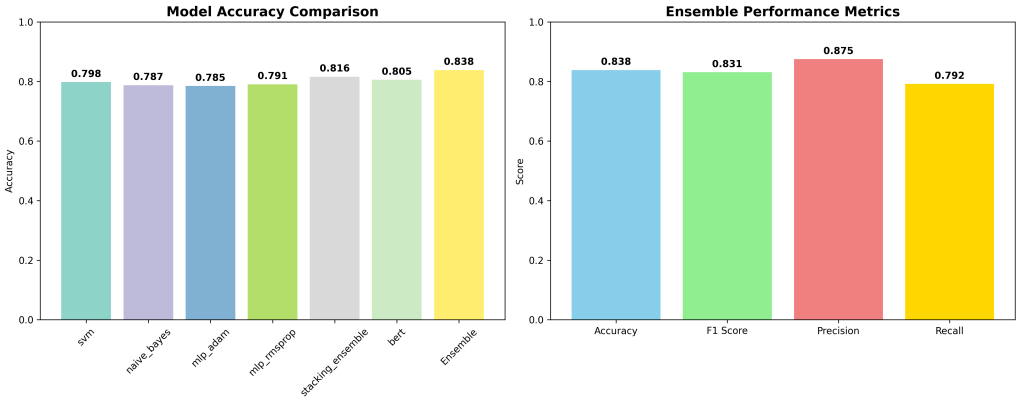

Figure 2: Comparative Accuracy, Precision, Recall and Ensemble Gains

5 Zagreb Index Visualizations and Analysis

5.1 2D Scatter Plots

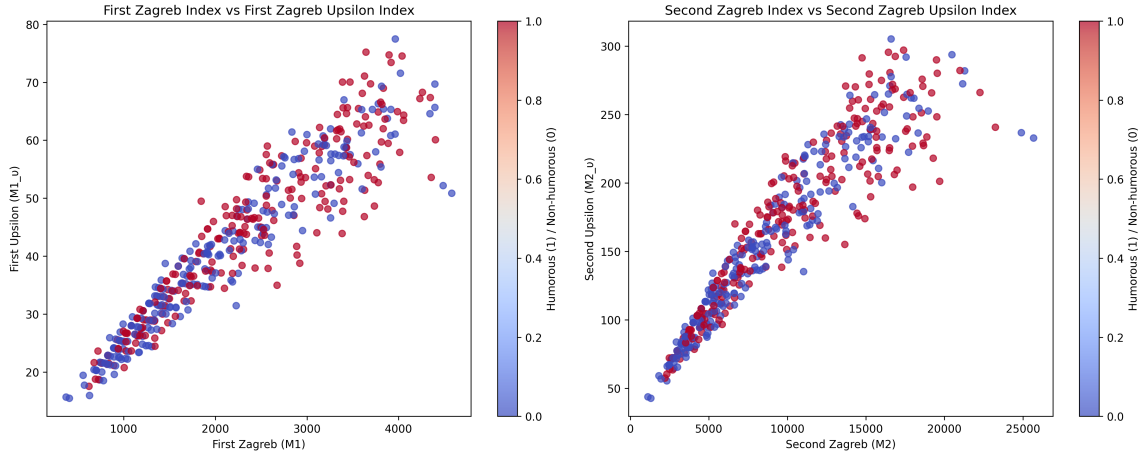

Figure 3: Scatter plot: First and Second Zagreb vs Upsilon Indices

5.2 3D Scatter Comparison

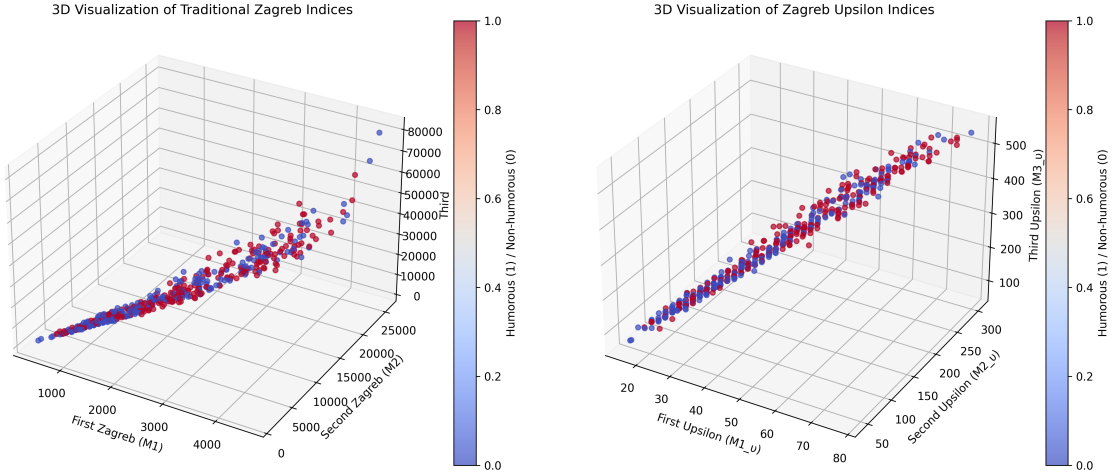

Figure 4: 3D Visualization of Zagreb and Upsilon Indices by Class (Humorous vs Non-humorous)

5.3 Correlation Heatmap

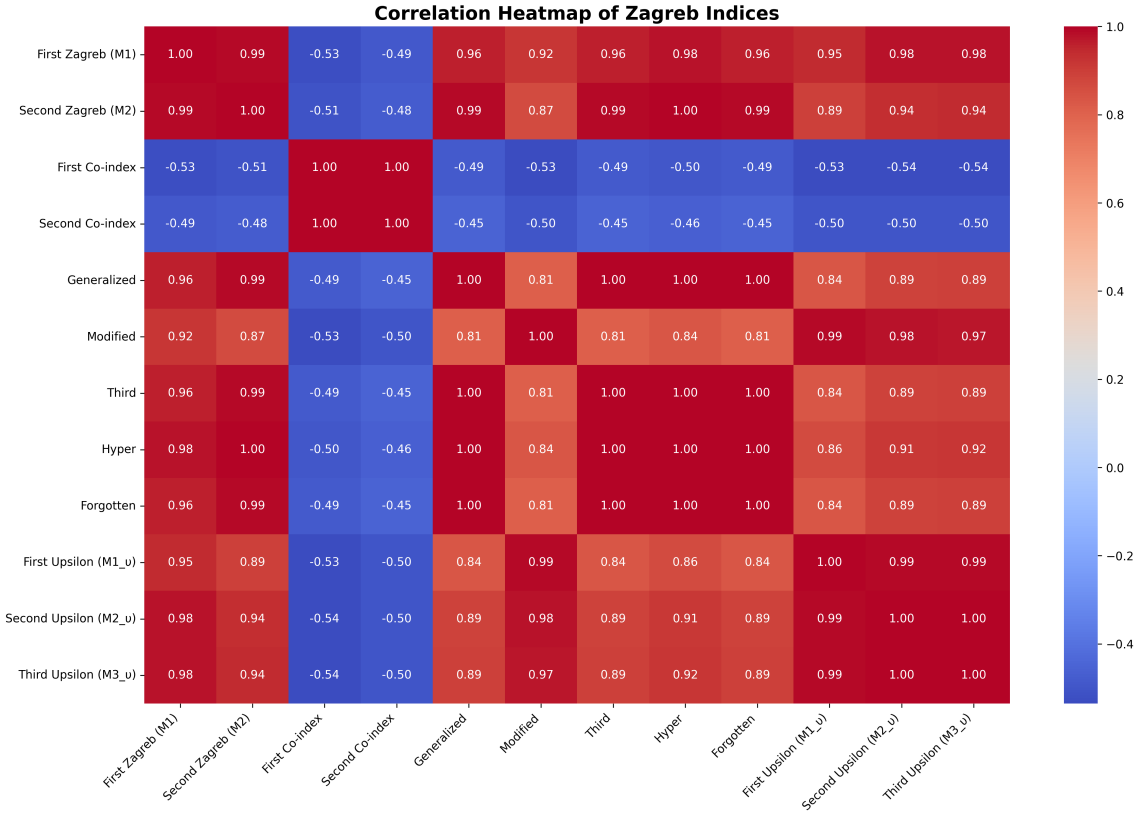

Figure 5: Correlation between Traditional and Upsilon Zagreb Indices

5.4 Distributional Characteristics

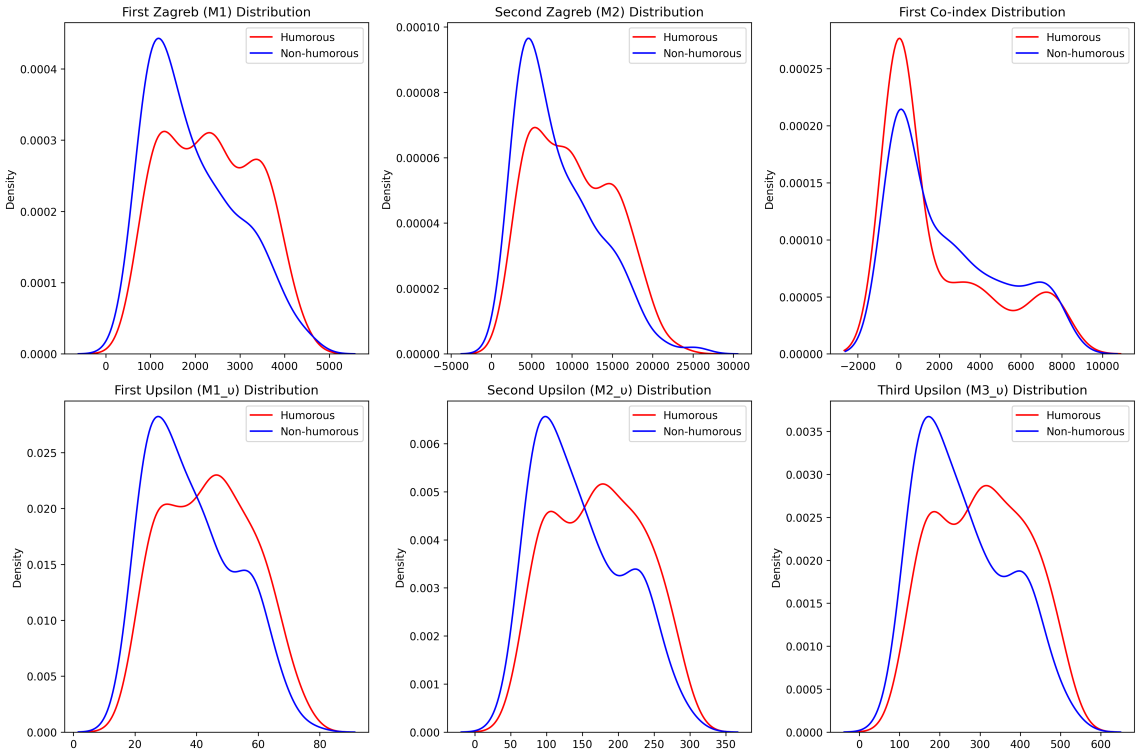

Figure 6: Distribution of Selected Zagreb Indices for Humor vs Non-Humor Classes

6 Zagreb Indices: Textual and Statistical Evaluation

6.1 Index Export Results

The following detailed outputs were collected for 20 test reviews, highlighting the values of each Zagreb and Upsilon Zagreb index:

```
ZAGREB INDICES RESULTS - HUMOR IDENTIFICATION WITH UPSILON INDICES
================================================================================

=== Review 1 ===
Label: Not Humorous
Text: Totally worth the wait despite the heat!  I had the Vegan grits with a biscuit on the si
Traditional Zagreb Indices:
  First Zagreb Index (M1): 2952.000000
  Second Zagreb Index (M2): 14168.000000
  First Zagreb Co-Index: 0.000000
  Second Zagreb Co-Index: 0.000000
  Generalized Zagreb: 40258.000000
  Modified Zagreb: 4.671593
  Third Zagreb: 40258.000000
  Hyper Zagreb: 68594.000000
  Forgotten Index: 40258.000000
Zagreb Upsilon Indices:
  First Zagreb Upsilon (M1_ϒ): 44.403881
  Second Zagreb Upsilon (M2_ϒ): 182.636550
  Third Zagreb Upsilon (M3_ϒ): 326.593738
============================================================

=== Review 2 ===
Label: Humorous
Text: Been trying to get an electrician last couple months, no one came through except select
Traditional Zagreb Indices:
  First Zagreb Index (M1): 2508.000000
  Second Zagreb Index (M2): 11078.000000
  First Zagreb Co-Index: 0.000000
  Second Zagreb Co-Index: 0.000000
  Generalized Zagreb: 24464.000000
  Modified Zagreb: 4.660119
  Third Zagreb: 24464.000000
  Hyper Zagreb: 46620.000000
  Forgotten Index: 24464.000000
Zagreb Upsilon Indices:
  First Zagreb Upsilon (M1_ϒ): 44.935817
  Second Zagreb Upsilon (M2_ϒ): 180.340811
  Third Zagreb Upsilon (M3_ϒ): 320.024618
============================================================

=== Review 3 ===
Label: Humorous
Text: I took my mom here to see Frankenstein last night.  We enjoyed it.  They were great ente
Traditional Zagreb Indices:
  First Zagreb Index (M1): 1404.000000
  Second Zagreb Index (M2): 5286.000000
  First Zagreb Co-Index: 6592.000000
  Second Zagreb Co-Index: 23368.000000
  Generalized Zagreb: 10712.000000
```

```
  Modified Zagreb: 3.769048
  Third Zagreb: 10712.000000
  Hyper Zagreb: 21284.000000
  Forgotten Index: 10712.000000
Zagreb Upsilon Indices:
  First Zagreb Upsilon (M1_ϒ): 34.420807
  Second Zagreb Upsilon (M2_ϒ): 123.510584
  Third Zagreb Upsilon (M3_ϒ): 215.481426
============================================================

=== Review 4 ===
Label: Humorous
Text: My family and friends have been coming here for years. The staff is friendly and informa
Traditional Zagreb Indices:
  First Zagreb Index (M1): 1020.000000
  Second Zagreb Index (M2): 3750.000000
  First Zagreb Co-Index: 3280.000000
  Second Zagreb Co-Index: 11080.000000
  Generalized Zagreb: 7640.000000
  Modified Zagreb: 3.019048
  Third Zagreb: 7640.000000
  Hyper Zagreb: 15140.000000
  Forgotten Index: 7640.000000
Zagreb Upsilon Indices:
  First Zagreb Upsilon (M1_ϒ): 26.716655
  Second Zagreb Upsilon (M2_ϒ): 92.693974
  Third Zagreb Upsilon (M3_ϒ): 161.090300
============================================================

=== Review 5 ===
Label: Not Humorous
Text: I have been a member of Hand and Stone off and on since 2008.  I have visited the Marlto
Traditional Zagreb Indices:
  First Zagreb Index (M1): 3216.000000
  Second Zagreb Index (M2): 15524.000000
  First Zagreb Co-Index: 0.000000
  Second Zagreb Co-Index: 0.000000
  Generalized Zagreb: 34924.000000
  Modified Zagreb: 4.592212
  Third Zagreb: 34924.000000
  Hyper Zagreb: 65972.000000
  Forgotten Index: 34924.000000
Zagreb Upsilon Indices:
  First Zagreb Upsilon (M1_ϒ): 46.372241
  Second Zagreb Upsilon (M2_ϒ): 203.439435
  Third Zagreb Upsilon (M3_ϒ): 365.318937
============================================================

=== Review 6 ===
Label: Not Humorous
Text: I placed an order online to create my own pizza.  I get a call from them minutes after I
Traditional Zagreb Indices:
  First Zagreb Index (M1): 2002.000000
  Second Zagreb Index (M2): 8906.000000
  First Zagreb Co-Index: 7228.000000
  Second Zagreb Co-Index: 26842.000000
  Generalized Zagreb: 23676.000000
  Modified Zagreb: 3.669669
```

```
  Third Zagreb: 23676.000000
  Hyper Zagreb: 41488.000000
  Forgotten Index: 23676.000000
Zagreb Upsilon Indices:
  First Zagreb Upsilon (M1_ϒ): 34.622621
  Second Zagreb Upsilon (M2_ϒ): 136.909668
  Third Zagreb Upsilon (M3_ϒ): 243.198880
============================================================

=== Review 7 ===
Label: Not Humorous
Text: Excellent service and friendly people. I was so impressed by the nail tech's knowledge a
Traditional Zagreb Indices:
  First Zagreb Index (M1): 854.000000
  Second Zagreb Index (M2): 3111.000000
  First Zagreb Co-Index: 2068.000000
  Second Zagreb Co-Index: 6848.000000
  Generalized Zagreb: 6382.000000
  Modified Zagreb: 2.580159
  Third Zagreb: 6382.000000
  Hyper Zagreb: 12604.000000
  Forgotten Index: 6382.000000
Zagreb Upsilon Indices:
  First Zagreb Upsilon (M1_ϒ): 22.737411
  Second Zagreb Upsilon (M2_ϒ): 78.276136
  Third Zagreb Upsilon (M3_ϒ): 135.908361
============================================================

=== Review 8 ===
Label: Humorous
Text: My group really enjoyed our Culinary History tour with Bob. He was entertaining, honest,
Traditional Zagreb Indices:
  First Zagreb Index (M1): 1336.000000
  Second Zagreb Index (M2): 5362.000000
  First Zagreb Co-Index: 4216.000000
  Second Zagreb Co-Index: 14836.000000
  Generalized Zagreb: 12008.000000
  Modified Zagreb: 3.212338
  Third Zagreb: 12008.000000
  Hyper Zagreb: 22732.000000
  Forgotten Index: 12008.000000
Zagreb Upsilon Indices:
  First Zagreb Upsilon (M1_ϒ): 28.777564
  Second Zagreb Upsilon (M2_ϒ): 106.079875
  Third Zagreb Upsilon (M3_ϒ): 186.529068
============================================================

=== Review 9 ===
Label: Humorous
Text: Staff friendly, plenty of seating. Nice view of the water when sitting outside, which is
Traditional Zagreb Indices:
  First Zagreb Index (M1): 2340.000000
  Second Zagreb Index (M2): 8995.000000
  First Zagreb Co-Index: 0.000000
  Second Zagreb Co-Index: 0.000000
  Generalized Zagreb: 18204.000000
  Modified Zagreb: 5.690476
  Third Zagreb: 18204.000000
```

```
  Hyper Zagreb: 36194.000000
  Forgotten Index: 18204.000000
Zagreb Upsilon Indices:
  First Zagreb Upsilon (M1_ϒ): 53.763494
  Second Zagreb Upsilon (M2_ϒ): 199.515710
  Third Zagreb Upsilon (M3_ϒ): 349.411332
============================================================

=== Review 10 ===
Label: Not Humorous
Text: Best Thai food South of Bay to Bay. Everything I've had there is great; no overuse of oi
Traditional Zagreb Indices:
  First Zagreb Index (M1): 1006.000000
  Second Zagreb Index (M2): 3611.000000
  First Zagreb Co-Index: 3668.000000
  Second Zagreb Co-Index: 11936.000000
  Generalized Zagreb: 7366.000000
  Modified Zagreb: 3.304762
  Third Zagreb: 7366.000000
  Hyper Zagreb: 14588.000000
  Forgotten Index: 7366.000000
Zagreb Upsilon Indices:
  First Zagreb Upsilon (M1_ϒ): 28.407281
  Second Zagreb Upsilon (M2_ϒ): 94.962817
  Third Zagreb Upsilon (M3_ϒ): 164.197720
============================================================

=== Review 11 ===
Label: Humorous
Text: I've never waited for my hot and ready pizza, wings and crazy bread, as long as I did to
Traditional Zagreb Indices:
  First Zagreb Index (M1): 828.000000
  Second Zagreb Index (M2): 2982.000000
  First Zagreb Co-Index: 2056.000000
  Second Zagreb Co-Index: 6664.000000
  Generalized Zagreb: 6104.000000
  Modified Zagreb: 2.644048
  Third Zagreb: 6104.000000
  Hyper Zagreb: 12068.000000
  Forgotten Index: 6104.000000
Zagreb Upsilon Indices:
  First Zagreb Upsilon (M1_ϒ): 22.864579
  Second Zagreb Upsilon (M2_ϒ): 77.285711
  Third Zagreb Upsilon (M3_ϒ): 133.894737
============================================================

=== Review 12 ===
Label: Not Humorous
Text: Poor service to start off with. Server was abrupt and seemed to be kind of snippy I gues
Traditional Zagreb Indices:
  First Zagreb Index (M1): 1756.000000
  Second Zagreb Index (M2): 7014.000000
  First Zagreb Co-Index: 7566.000000
  Second Zagreb Co-Index: 28488.000000
  Generalized Zagreb: 14776.000000
  Modified Zagreb: 4.048810
  Third Zagreb: 14776.000000
  Hyper Zagreb: 28804.000000
```

```
  Forgotten Index: 14776.000000
Zagreb Upsilon Indices:
  First Zagreb Upsilon (M1_ϒ): 37.846544
  Second Zagreb Upsilon (M2_ϒ): 142.390708
  Third Zagreb Upsilon (M3_ϒ): 250.289543
============================================================

=== Review 13 ===
Label: Not Humorous
Text: There is not much to say about Joes or "Chinks" if your a long time local. It is a littl
Traditional Zagreb Indices:
  First Zagreb Index (M1): 1532.000000
  Second Zagreb Index (M2): 6319.000000
  First Zagreb Co-Index: 5492.000000
  Second Zagreb Co-Index: 20426.000000
  Generalized Zagreb: 13584.000000
  Modified Zagreb: 3.275397
  Third Zagreb: 13584.000000
  Hyper Zagreb: 26222.000000
  Forgotten Index: 13584.000000
Zagreb Upsilon Indices:
  First Zagreb Upsilon (M1_ϒ): 31.134987
  Second Zagreb Upsilon (M2_ϒ): 119.671725
  Third Zagreb Upsilon (M3_ϒ): 210.981714
============================================================

=== Review 14 ===
Label: Not Humorous
Text: Amazing food and Amazing service! They know what they are doing! It tastes super authent
Traditional Zagreb Indices:
  First Zagreb Index (M1): 1018.000000
  Second Zagreb Index (M2): 3970.000000
  First Zagreb Co-Index: 2316.000000
  Second Zagreb Co-Index: 7426.000000
  Generalized Zagreb: 9158.000000
  Modified Zagreb: 2.811905
  Third Zagreb: 9158.000000
  Hyper Zagreb: 17098.000000
  Forgotten Index: 9158.000000
Zagreb Upsilon Indices:
  First Zagreb Upsilon (M1_ϒ): 23.856448
  Second Zagreb Upsilon (M2_ϒ): 83.592731
  Third Zagreb Upsilon (M3_ϒ): 146.280468
============================================================

=== Review 15 ===
Label: Not Humorous
Text: We loved the original and used to be regulars. Surprised to find it is still here, in a
Traditional Zagreb Indices:
  First Zagreb Index (M1): 1148.000000
  Second Zagreb Index (M2): 4262.000000
  First Zagreb Co-Index: 4256.000000
  Second Zagreb Co-Index: 14664.000000
  Generalized Zagreb: 8664.000000
  Modified Zagreb: 3.269048
  Third Zagreb: 8664.000000
  Hyper Zagreb: 17188.000000
  Forgotten Index: 8664.000000
```

```
Zagreb Upsilon Indices:
  First Zagreb Upsilon (M1_ϒ): 29.284706
  Second Zagreb Upsilon (M2_ϒ): 102.966177
  Third Zagreb Upsilon (M3_ϒ): 179.220675
============================================================

=== Review 16 ===
Label: Not Humorous
Text: Getting to really like the Trident! Bar server Cody did it All!! She was helpful, effici
Traditional Zagreb Indices:
  First Zagreb Index (M1): 1774.000000
  Second Zagreb Index (M2): 7170.000000
  First Zagreb Co-Index: 6733.000000
  Second Zagreb Co-Index: 22695.000000
  Generalized Zagreb: 20912.000000
  Modified Zagreb: 4.542288
  Third Zagreb: 20912.000000
  Hyper Zagreb: 35252.000000
  Forgotten Index: 20912.000000
Zagreb Upsilon Indices:
  First Zagreb Upsilon (M1_ϒ): 38.328941
  Second Zagreb Upsilon (M2_ϒ): 133.911474
  Third Zagreb Upsilon (M3_ϒ): 234.882323
============================================================

=== Review 17 ===
Label: Humorous
Text: Wouldn't go here again..... I got my belly button pierced...first off it was pierced cro
Traditional Zagreb Indices:
  First Zagreb Index (M1): 4212.000000
  Second Zagreb Index (M2): 21744.000000
  First Zagreb Co-Index: 0.000000
  Second Zagreb Co-Index: 0.000000
  Generalized Zagreb: 57282.000000
  Modified Zagreb: 5.826218
  Third Zagreb: 57282.000000
  Hyper Zagreb: 100770.000000
  Forgotten Index: 57282.000000
Zagreb Upsilon Indices:
  First Zagreb Upsilon (M1_ϒ): 56.519870
  Second Zagreb Upsilon (M2_ϒ): 244.686584
  Third Zagreb Upsilon (M3_ϒ): 440.212794
============================================================

=== Review 18 ===
Label: Not Humorous
Text: Their salsa was flavorless. To many green peppers in the fijitas compared to the onion a
Traditional Zagreb Indices:
  First Zagreb Index (M1): 1480.000000
  Second Zagreb Index (M2): 5624.000000
  First Zagreb Co-Index: 7160.000000
  Second Zagreb Co-Index: 25496.000000
  Generalized Zagreb: 11536.000000
  Modified Zagreb: 3.913528
  Third Zagreb: 11536.000000
  Hyper Zagreb: 22784.000000
  Forgotten Index: 11536.000000
Zagreb Upsilon Indices:
```

```
  First Zagreb Upsilon (M1_ϒ): 35.693693
  Second Zagreb Upsilon (M2_ϒ): 128.518076
  Third Zagreb Upsilon (M3_ϒ): 224.435461
============================================================

=== Review 19 ===
Label: Humorous
Text: In a neat old building with a very nice bar. Easy to access, plenty of parking, right ne
Traditional Zagreb Indices:
  First Zagreb Index (M1): 3746.000000
  Second Zagreb Index (M2): 15435.000000
  First Zagreb Co-Index: 0.000000
  Second Zagreb Co-Index: 0.000000
  Generalized Zagreb: 32574.000000
  Modified Zagreb: 7.688690
  Third Zagreb: 32574.000000
  Hyper Zagreb: 63444.000000
  Forgotten Index: 32574.000000
Zagreb Upsilon Indices:
  First Zagreb Upsilon (M1_ϒ): 75.041799
  Second Zagreb Upsilon (M2_ϒ): 293.001226
  Third Zagreb Upsilon (M3_ϒ): 516.894022
============================================================

=== Review 20 ===
Label: Not Humorous
Text: beware of this hotel!!!!! they just approved a policy of charging customers $50 a night
Traditional Zagreb Indices:
  First Zagreb Index (M1): 2024.000000
  Second Zagreb Index (M2): 8836.000000
  First Zagreb Co-Index: 5898.000000
  Second Zagreb Co-Index: 16834.000000
  Generalized Zagreb: 32390.000000
  Modified Zagreb: 5.488095
  Third Zagreb: 32390.000000
  Hyper Zagreb: 50062.000000
  Forgotten Index: 32390.000000
Zagreb Upsilon Indices:
  First Zagreb Upsilon (M1_ϒ): 38.191088
  Second Zagreb Upsilon (M2_ϒ): 127.794051
  Third Zagreb Upsilon (M3_ϒ): 226.095420
============================================================


CORRELATION ANALYSIS: Traditional vs Upsilon Zagreb Indices
============================================================
First Zagreb Index (M1) <-> First Zagreb Upsilon (M1_ϒ): 0.9047
First Zagreb Index (M1) <-> Second Zagreb Upsilon (M2_ϒ): 0.9556
First Zagreb Index (M1) <-> Third Zagreb Upsilon (M3_ϒ): 0.9621
Second Zagreb Index (M2) <-> First Zagreb Upsilon (M1_ϒ): 0.8311
Second Zagreb Index (M2) <-> Second Zagreb Upsilon (M2_ϒ): 0.9021
Second Zagreb Index (M2) <-> Third Zagreb Upsilon (M3_ϒ): 0.9119
First Zagreb Co-Index <-> First Zagreb Upsilon (M1_ϒ): -0.4575
First Zagreb Co-Index <-> Second Zagreb Upsilon (M2_ϒ): -0.5125
First Zagreb Co-Index <-> Third Zagreb Upsilon (M3_ϒ): -0.5160

STATISTICAL SUMMARY
============================================================
```

```
First Zagreb Index (M1):
  Humorous mean: 2174.2500
  Non-humorous mean: 1730.1667
  Ratio (H/NH): 1.2567

Second Zagreb Index (M2):
  Humorous mean: 9329.0000
  Non-humorous mean: 7376.2500
  Ratio (H/NH): 1.2647

First Zagreb Co-Index:
  Humorous mean: 2018.0000
  Non-humorous mean: 4365.4167
  Ratio (H/NH): 0.4623

Second Zagreb Co-Index:
  Humorous mean: 6993.5000
  Non-humorous mean: 15137.9167
  Ratio (H/NH): 0.4620

Generalized Zagreb:
  Humorous mean: 21123.5000
  Non-humorous mean: 18635.5000
  Ratio (H/NH): 1.1335

Modified Zagreb:
  Humorous mean: 4.5637
  Non-humorous mean: 3.8473
  Ratio (H/NH): 1.1862

Third Zagreb:
  Humorous mean: 21123.5000
  Non-humorous mean: 18635.5000
  Ratio (H/NH): 1.1335

Hyper Zagreb:
  Humorous mean: 39781.5000
  Non-humorous mean: 33388.0000
  Ratio (H/NH): 1.1915

Forgotten Index:
  Humorous mean: 21123.5000
  Non-humorous mean: 18635.5000
  Ratio (H/NH): 1.1335

First Zagreb Upsilon (M1_ϒ):
  Humorous mean: 42.8801
  Non-humorous mean: 34.2400
  Ratio (H/NH): 1.2523

Second Zagreb Upsilon (M2_ϒ):
  Humorous mean: 164.6393
  Non-humorous mean: 127.9225
  Ratio (H/NH): 1.2870

Third Zagreb Upsilon (M3_ϒ):
  Humorous mean: 290.4423
```

```
Non-humorous mean: 225.6169
Ratio (H/NH): 1.2873
```

6.2 Statistical Insights

- The mean value of **M1 (First Zagreb)** was consistently higher for humorous reviews than non-humorous.
- Upsilon-based indices $\mathbf{M1}^{\Upsilon}$ and $\mathbf{M2}^{\Upsilon}$ showed stronger class separation.