Yelp Reviews Clustering: NLP & Unsupervised ML

Johnny Ye & Xiuyun Wen

Outline

1. **Background**
2. **Motivation**
3. **Methodology**
4. **Result**

Background

Literature Review:

- **Data**: labeled data (no source), manually labeled data
- **Features**: extreme ratings, semantic models, shorter in length, users have fewer friends and a lower review count, fake reviews had a slightly higher average rating than real reviews, other information (user’s account activity etc).
- **Model**: supervised learning. Naïve Bayes, SVM, and Decision Tree, RF, NN, logistic regression

Motivation

Proposed innovative method:

- **Goal:** using unsupervised clustering method on unlabeled data to identify suspicious/unauthentic reviews.
- **Data**: not labeled data.
- **Features**:

  extreme ratings/higher than average

  Sentiment score ; Subjectivity

  shorter in length

  fewer friends (how many people agree)

Methodology

Data: 10001 rows / from Yelp.com

Variables:

Subjectivity Score:

- Float, ranges from 0 to 1. 0 indicates objective;
- 1 suggests subjective.
- Get from TextBlob Model.

Word_count:

- Number of review words.
- Integer

FCU_count:

- the number of people think a review is funny/cool/useful.
- Integer

Absolute_diff:

- the absolute difference between review stars and polarity scores
- Float

Abosulte_diff_x_p:

- the absolute difference between restaurant stars and polarity scores
- Float

Text Embedding:

- Feature extraction by CLIP Model
- Float

Methodology

Method

NLP:
1. RoBERTa Model: Polarity Score
2. Textblob: Subjectivity Score
3. CLIP: Text Feature Extraction

Unsupervised ML:

Dimensionality Reduction
1. PCA
2. TSNE

Clustering
1. K-Means
2. Hierarchical Clustering

Evaluation Method

1. Silhouette Score: Evaluate clustering

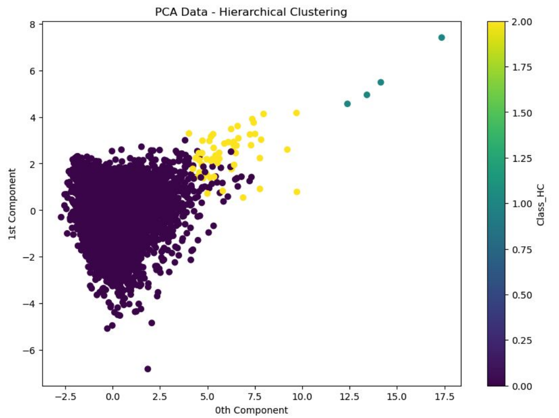
2. Davies-Bouldin Index: Evaluate clustering

3. Cohen's Kappa: Compare the label results between made from model and manually

Result

1. Silhouette Score & Davies–Bouldin Index

	Metric	Value
0	Silhouette Score (PCA KMeans)	0.231627
1	Davies-Bouldin Index (PCA KMeans)	1.530860
2	Silhouette Score (PCA Hierarchical)	0.662360
3	Davies-Bouldin Index (PCA Hierarchical)	0.498707
4	Silhouette Score (t-SNE KMeans, CLIP)	0.437561
5	Davies-Bouldin Index (t-SNE KMeans, CLIP)	0.763920
6	Silhouette Score (t-SNE Hierarchical, CLIP)	0.334346
7	Davies-Bouldin Index (t-SNE Hierarchical, CLIP)	0.751416



Silhouette Score: 1: well matched to their own cluster and poorly matched to neighboring clusters.
Davies–Bouldin Index: evaluates the clustering quality by considering the ratio of within-cluster distances to between-cluster distances. The lower the Index, the better the clustering.

Result

2. Metrics

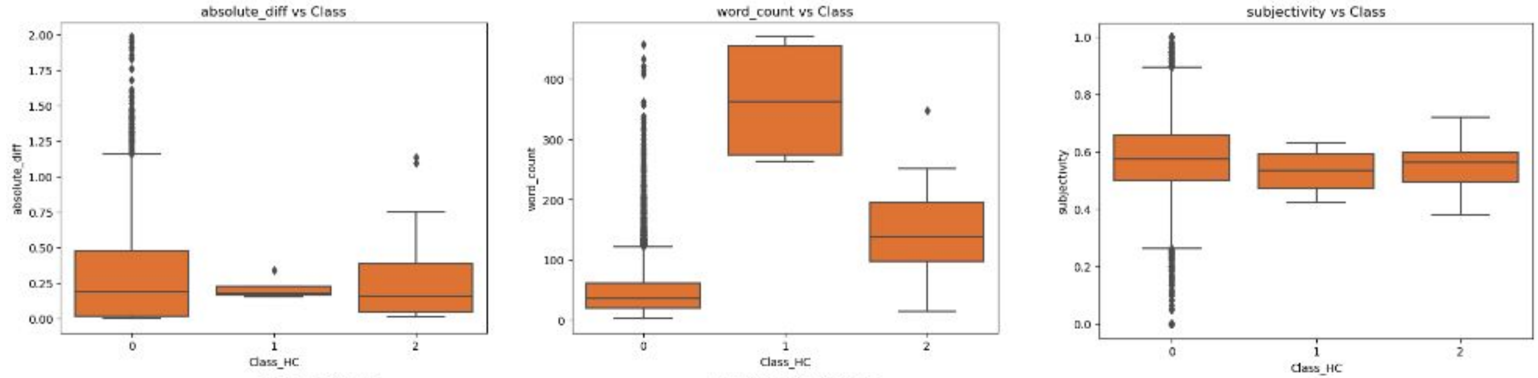
Manually labeled data: 200 reviews with 0,1,2 cluster.

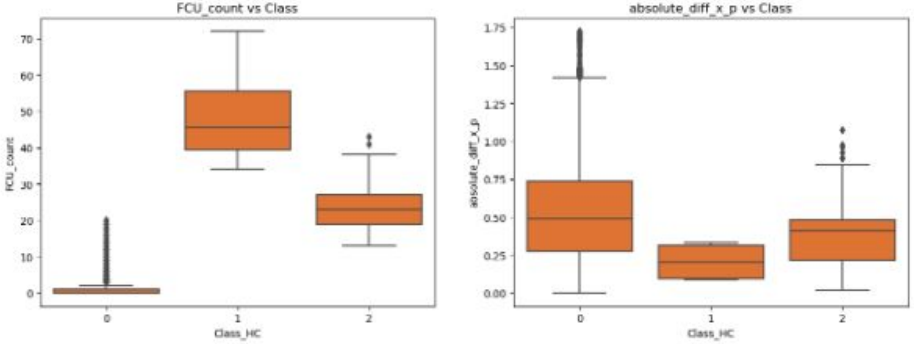
Cohen's kappa coefficient: slight level of agreement (0.1-0.2)

```
# Calculate Cohen's Kappa Score
cohen_kappa = cohen_kappa_score(label['Class_HC'], label['label'])
print(cohen_kappa)
✓ 0.0s

0.13967611336032393
```

Result

2.





1: Useful, informative
2: Valid,normal comments
0: Suspicious or not useful

References

Kossakov, M., Mukasheva, A., Balbayev, G., Seidazimov, S., Mukammejanova, D., & Sydybayeva, M. (2024). Quantitative comparison of machine learning clustering methods for tuberculosis data analysis. CIEES 2023. https://doi.org/10.3390/engproc2024060020

Li, Y., Feng, X., & Zhang, S. (2016). Detecting Fake Reviews Utilizing Semantic and Emotion Model. In 2016 3rd International Conference on Information Science and Control Engineering (ICISCE). IEEE. https://doi.org/10.1109/ICISCE.2016.77

Luca, M., & Zervas, G. (2016). Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud. Management Science, 62(12), 3412-3427. https://doi.org/10.1287/mnsc.2015.2304;

Richards, J., Dabhi, S., Poursardar, F., & Jayarathna, S. (2023). Poster: Leveraging Data Analysis and Machine Learning to Authenticate Yelp Reviews. ACM, New York, NY, USA. https://doi.org/10.1145/3565287.3617983

Thanks

Sentiment Analysis - RoBERTa

- **Contextual Understanding**:

  trained on vast amounts of text data. Understand language in its natural form.

- **Pre-trained Tokenization**:

  RoBERTa comes with its own tokenizer. used by transformer-based models, splitting the text into tokens in a way that's optimal for the model.

- **Handling of Special Tokens**:

  adds special tokens for the model to understand the structure of the text. For example, help the model understand the beginning and end of a text segment.

Features selection

- **Subjectivity Score:**
  TextBlob. ranges from 0 to 1. 0 indicates objective; 1 suggests subjective.
- **Word_count:**
  the number of words in processed reviews.
- **FCU_count:**
  the number of people think a review is funny/cool/useful.
- **Absolute_diff:**
  the absolute difference between y_stars (the star rating given by the user in their review) and **roberta_polarity** (calculated from the reviews using pre-trained RoBERTa model)
- **Abosulte_diff_x_p:**
  the absolute difference between x_stars (the average star rating of the restaurant.) and roberta_polarity.
- **CLIP model:**
  Text embedding. Txt -> high dim vector. Use as feature extraction.

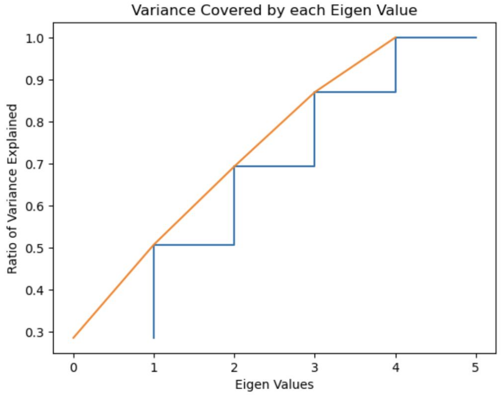
Unsupervised ML Strategy

1. Use Yelp data & NLP features **(Subjectivity Score, Absolute_diff, Word_count, FCU_count, Abosulte_diff_x_p:)**
    1.1. Use PCA to do dimensionality reduction
        1.1.1. K-Means
        1.1.2. Hierarchical Clustering
2. Use Yelp data & NLP features + features from CLIP model + TSNE
    2.1. Add features from CLIP model
    2.2. Use TSNE to do dimensionality reduction
        2.2.1. Fine-tuning TSNE hyperparameter
        2.2.2. K-Means
        2.2.3. Hierarchical Clustering
3. Results Comparison -> Optimal model
    3.1. Compare the silhouette score and Davies-Bouldin Index of models
4. Use optimal model to plot statistical graphs and get qualitative scores

PCA - Dimensionality Reduction

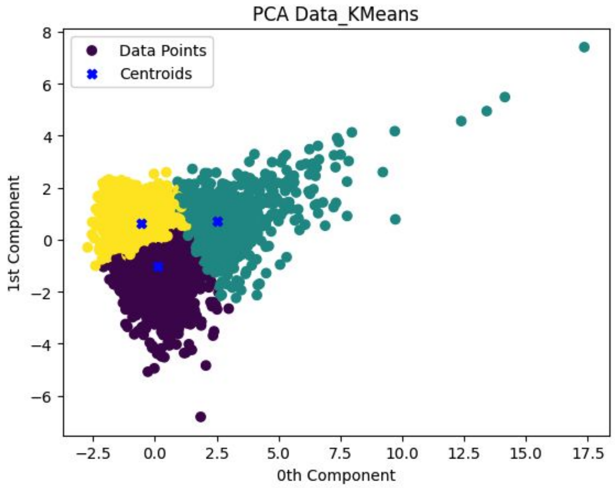
Standardization

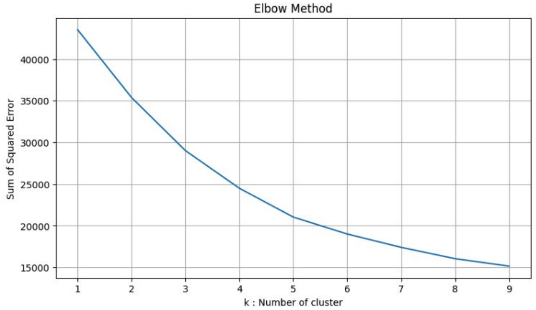
	absolute_diff	word_count	FCU_count	absolute_diff_x_p	subjectivity
0	1.217792	-0.152997	-0.438616	0.310831	-1.252583
1	2.217123	-0.266684	-0.438616	0.477548	-0.418638
2	-0.885254	-0.585006	0.242427	-0.213499	1.182624
3	0.639675	0.256274	0.242427	-0.226694	-0.661121
4	-0.611694	-0.334896	0.923470	2.581931	-1.302281
...	...	...	...	...	...
9996	-0.856209	-0.789641	-0.438616	-0.240504	-0.965863
9997	-0.905622	0.483646	-0.098094	-0.194561	0.149413
9998	-0.907669	-0.289421	-0.438616	-0.955103	0.484939
9999	5.150970	-0.675955	-0.438616	0.155612	0.396629
10000	-0.814787	0.165324	-0.438616	-0.279019	-0.307483

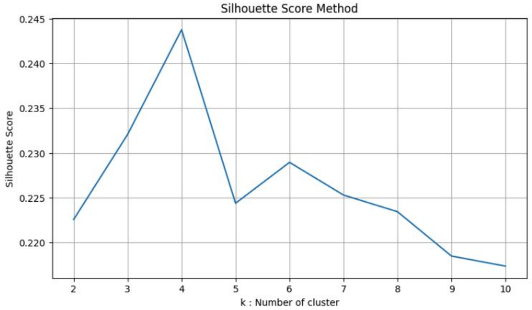


	0	1	2	3
0	0.410348	-1.541601	-0.867482	-0.141160
1	0.230728	-1.958033	-0.685803	1.081904
2	-0.912232	1.107123	0.668927	0.248585
3	0.757119	-0.402950	-0.484047	0.060063
4	0.366355	-1.586993	1.886469	-1.779163
...	...	...	...	...
9996	-0.572359	0.103190	-0.603422	-1.282764
9997	0.003060	0.771931	0.163751	-0.463353
9998	-0.746467	1.203567	-0.454310	-0.144999
9999	0.400550	-3.287948	-1.422337	3.702821
10000	-0.188190	0.473985	-0.299876	-0.745249

K-Means - PCA



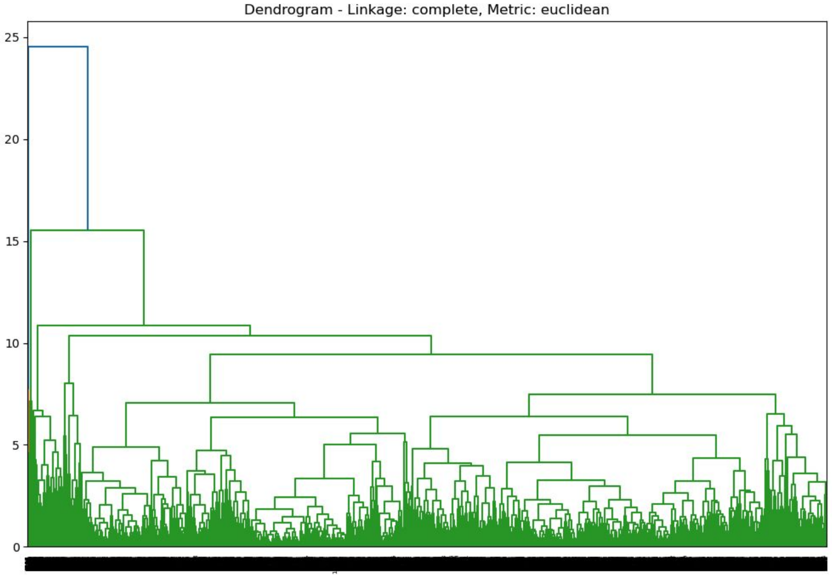


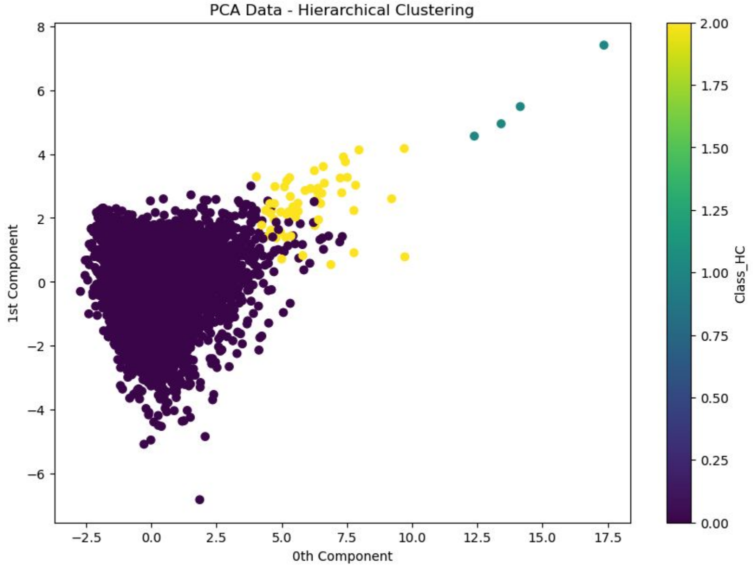


silhouette_score_PCA_KMEANS: 0.2316270373287704 1 is the best

Davies-Bouldin Index PCA_KEAMNS: 1.5308598600091132 Lower is better

Hierarchical Clustering -PCA






```
silhouette_score_PCA_Hier: 0.662360352537421
Davies-Bouldin Index PCA_Hier: 0.4987070102042927
```

1 is the best
Lower is better

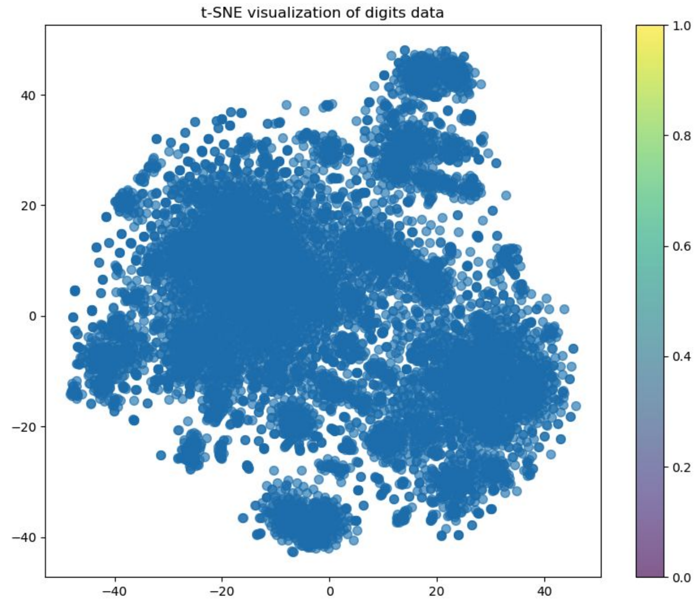
Data + CLIP

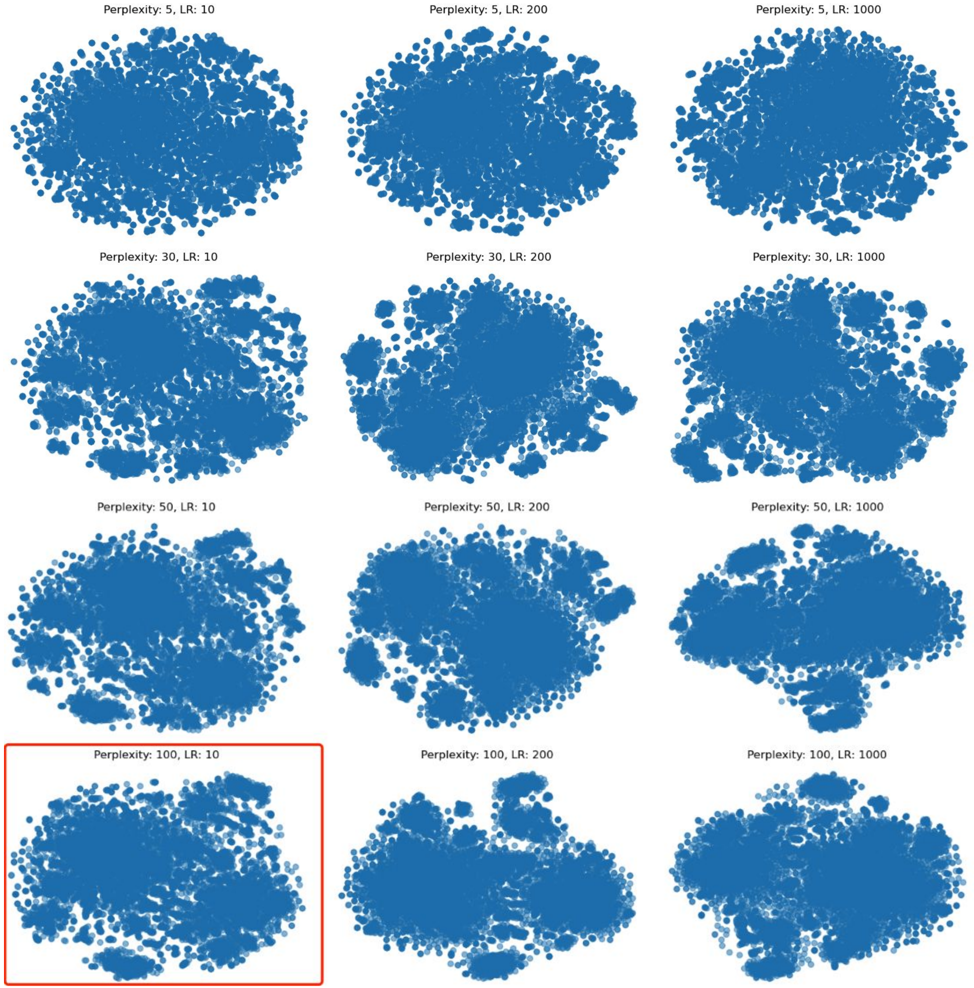
	subjectivity	absolute_diff	word_count	FCU_count	absolute_diff_x_p	embed_0	embed_1	embed_2
0	0.412121	0.656539	42	0.0	0.656539	-0.073401	-0.041123	0.012648
1	0.522294	0.961204	37	0.0	0.711204	0.186031	0.135670	0.047222
2	0.733838	0.015385	23	2.0	0.484615	0.187688	0.079294	-0.159329
3	0.490260	0.480289	60	2.0	0.480289	0.051479	0.052152	-0.128900
4	0.405556	0.098785	34	4.0	1.401215	0.064015	0.070755	0.067738
...	...	...	...	...	...	...	...	...
9996	0.450000	0.024239	14	0.0	0.475761	0.115848	-0.195240	0.141284
9997	0.597340	0.009175	70	1.0	0.490825	0.348834	-0.290700	0.071015
9998	0.641667	0.008551	36	0.0	0.241449	0.338024	-0.116429	0.003233
9999	0.630000	1.855644	19	0.0	0.605644	-0.001474	0.170802	-0.043595
10000	0.536979	0.036868	56	0.0	0.463132	0.177886	-0.310400	0.038181

10001 rows × 517 columns

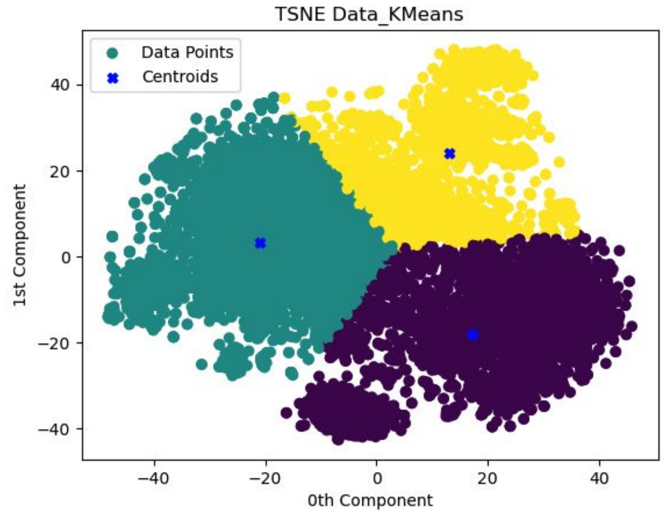
TSNE

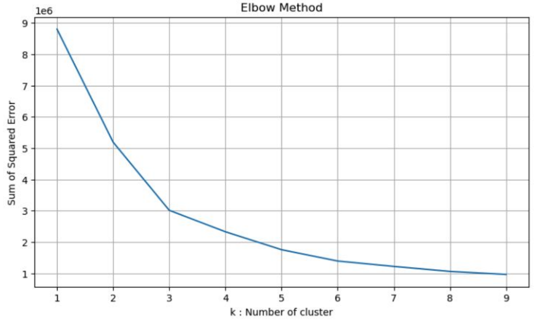
```
tsne = TSNE(n_components=2, random_state=42, perplexity=100, learning_rate=10, n_iter=10000)

df_tsne = tsne.fit_transform(final_dataframe_CLIP)

# plot
plt.figure(figsize=(10, 8))
plt.scatter(df_tsne[:, 0], df_tsne[:, 1], cmap='viridis', s=50, alpha=0.6)
plt.colorbar()
plt.title('t-SNE visualization of digits data')
plt.show()
```

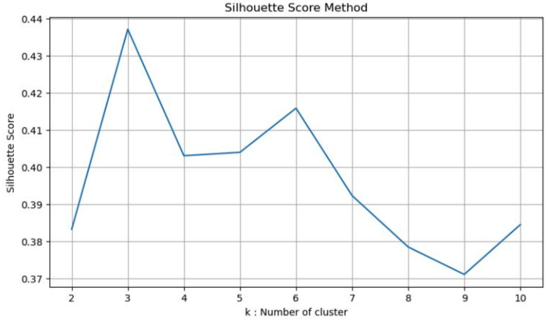




K-Means - TSNE



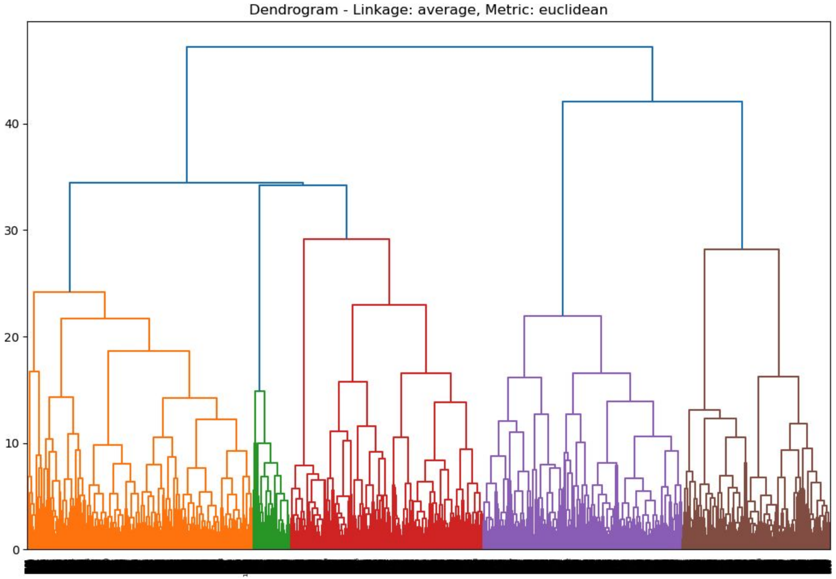


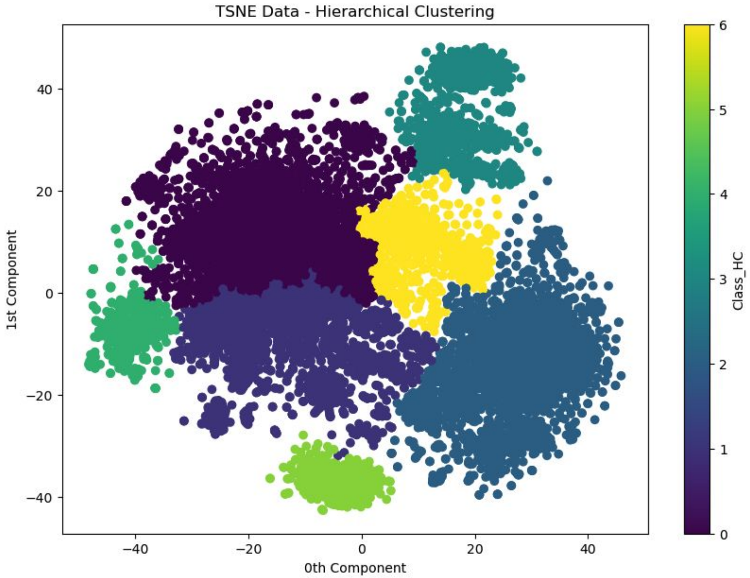


silhouette_score_TSNE_KMEANS_CLIP: 0.4375605285167694
Davies-Bouldin Index TSNE_KEAMNS_CLIP: 0.7639196816156874

1 is the best
Lower is better

Hierarchical Clustering - TSNE





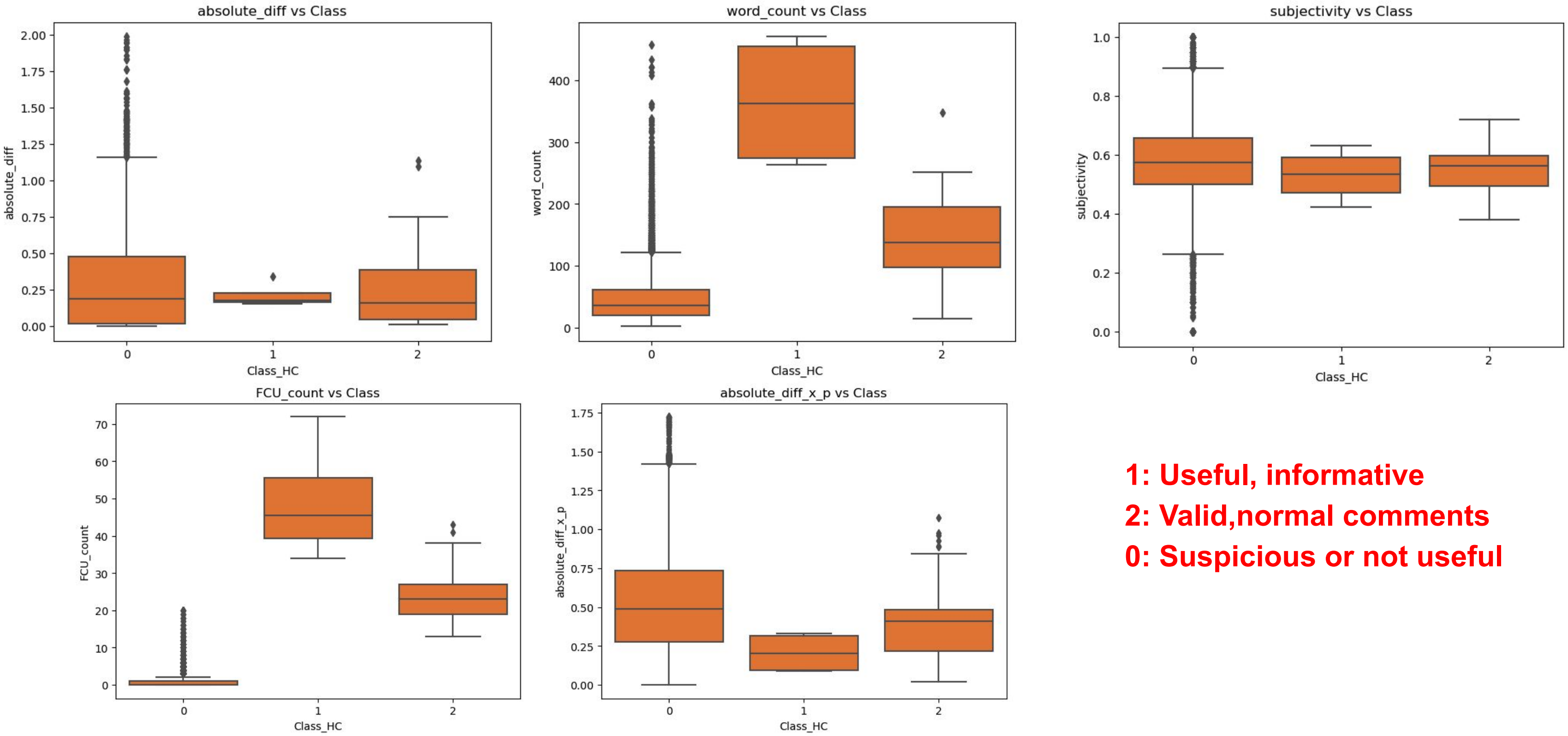
silhouette_score_TSNE_Hier: 0.3343462646007538
Davies-Bouldin Index TSNE_Hier: 0.7514161057220408

1 is the best
Lower is better

Model Results Comparison

	Metric	Value
0	Silhouette Score (PCA KMeans)	0.231627
1	Davies-Bouldin Index (PCA KMeans)	1.530860
2	Silhouette Score (PCA Hierarchical)	0.662360
3	Davies-Bouldin Index (PCA Hierarchical)	0.498707
4	Silhouette Score (t-SNE KMeans, CLIP)	0.437561
5	Davies-Bouldin Index (t-SNE KMeans, CLIP)	0.763920
6	Silhouette Score (t-SNE Hierarchical, CLIP)	0.334346
7	Davies-Bouldin Index (t-SNE Hierarchical, CLIP)	0.751416

Statistical Plots & Qualitative Score (PCA Hierarchical - optimal model)



1: Useful, informative
2: Valid,normal comments
0: Suspicious or not useful

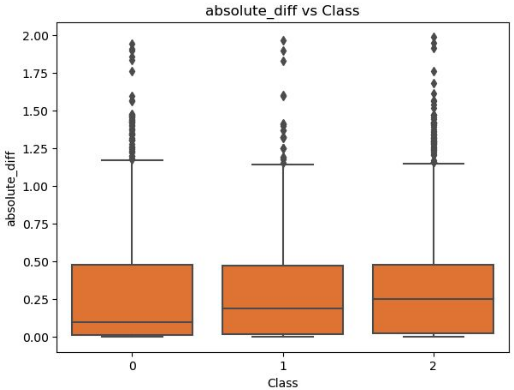
Conclusion

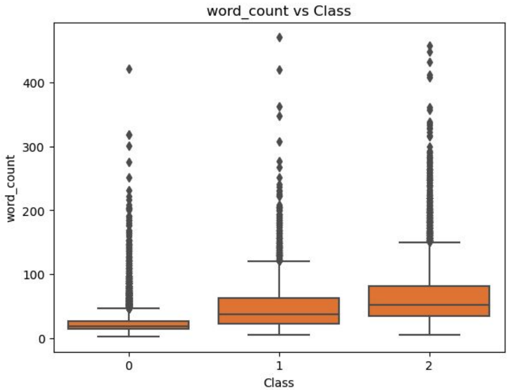
Unsupervised learning with features inspired by literature reviews.

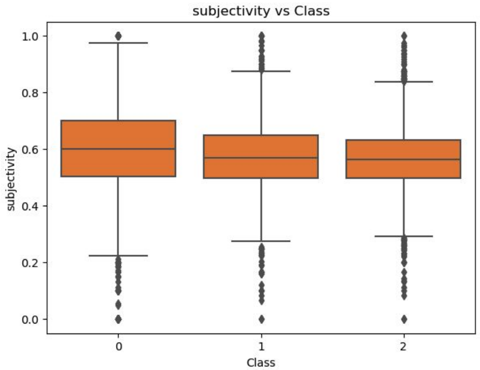
Silhouette Score(0.66), Davies–Bouldin Index(0.49), Cohen's kappa (0.13).

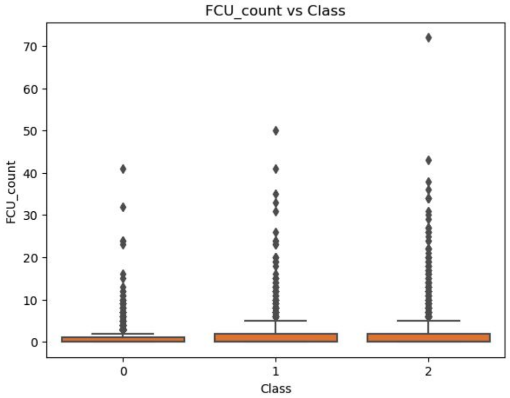
The above model is able to generate a good cluster, but only slightly agree with our labeled result. Model has room for improvement. (semi-supervised learning, better features etc)

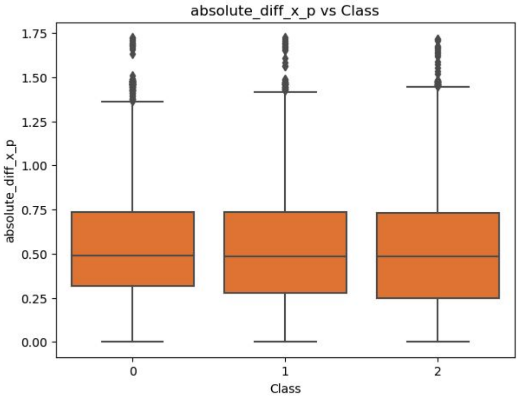
Statistical Plots & Qualitative Score (TSNE+CLIP+KMeans)











- **Differences are almost the same**
- **Slight gaps in text length and text subjectivity**

Summary

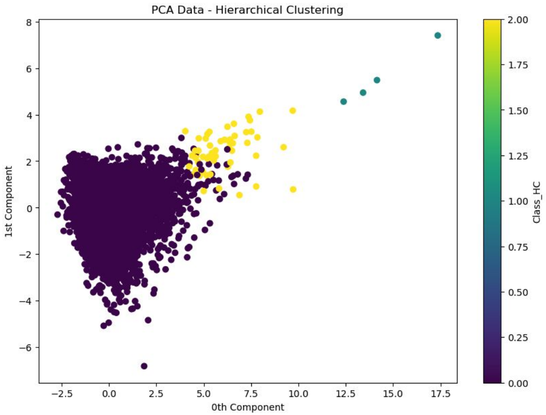
NLP

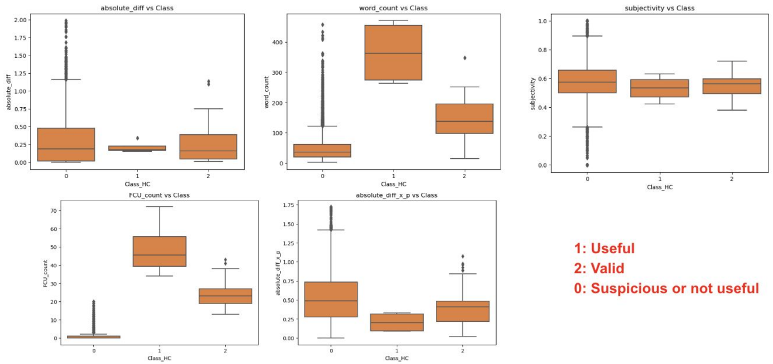
Roberta polarity;
Subjectivity

Features Selection

Subjectivity Score;
Absolute_diff;
Word_count;
FCU_count;
Abosulte_diff_x_p;
Feature Extraction from CLIP model

Unsupervised ML





1: Useful
2: Valid
0: Suspicious or not useful

References

Kossakov, M., Mukasheva, A., Balbayev, G., Seidazimov, S., Mukammejanova, D., & Sydybayeva, M. (2024). Quantitative comparison of machine learning clustering methods for tuberculosis data analysis. CIEES 2023. https://doi.org/10.3390/engproc2024060020

Li, Y., Feng, X., & Zhang, S. (2016). Detecting Fake Reviews Utilizing Semantic and Emotion Model. In 2016 3rd International Conference on Information Science and Control Engineering (ICISCE). IEEE. https://doi.org/10.1109/ICISCE.2016.77

Luca, M., & Zervas, G. (2016). Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud. Management Science, 62(12), 3412-3427. https://doi.org/10.1287/mnsc.2015.2304;

Richards, J., Dabhi, S., Poursardar, F., & Jayarathna, S. (2023). Poster: Leveraging Data Analysis and Machine Learning to Authenticate Yelp Reviews. ACM, New York, NY, USA. https://doi.org/10.1145/3565287.3617983

Thanks