An Unsupervised Analysis of Linguistic Personas in Movie Reviews

Brandon Baek
Crescenta Valley High School
La Crescenta, USA
brandonbaek2010@gmail.com

***Abstract*—This research investigates an unsupervised approach to analyzing distinct linguistic personas in movie reviews, focusing on writing style. By clustering reviews using stylometric features such as sentence complexity and vocabulary richness, this study identifies and analyzes unique linguistic personalities within movie reviews. The resulting system provides a deeper understanding of opinion diversity, enabling more nuanced interpretations of user feedback beyond simple polarity.**

***Index Terms*—linguistic analysis, unsupervised learning, clustering, stylometry, movie reviews, persona analysis**

I. Introduction

Online reviews have become a cornerstone of modern decision-making. Yet, not all reviewers express themselves in the same way; some write emotionally charged rants, while others provide detailed, analytical feedback. This diversity in writing style holds untapped potential for deeper insights. This research investigates how linguistic patterns can uncover distinct reviewer personas and examines how the behavior of these personas differs in their thematic focus, sentiment, and more. This offers a greater understanding of public opinion.

II. Literature Review

Finding specific linguistic cues to analyze patterns is pivotal to this study. A golden standard for these text features is Pennebaker's LIWC feature [1], whose quality is evident in the work of Mairesse et al. [2] and Celli [3].

Mairesse et al.'s work is foundational, experimenting with recognizing Big 5 personality traits using various feature sets and proving lexical relations to character. [2].

Celli's work digs deeper, filtering Mairesse et al.'s features to identify the 22 most correlated ones. Celli explores using these features to predict personality even in short online texts [3].

Although these works are crucial, they explore personality recognition. My work aims to recognize writing style. Personality may correlate with writing style, but they are distinct concepts.

Sharma et al.'s work shares many similarities with the goals of my research [4]. They use clustering analysis on the characteristics of a review. However, they use a dataset that gives them access to data about each reviewer, giving them the ability to see the total amount of reviews they have given in the past, location, gender, age, and a lot more. Sharma et al. divides reviewer helpfulness into 4 different factors: Expertise, Experience, Identity Disclosure, and Review Characteristics. For review characteristics, they only identify 3 features. [4] My study will be solely focusing on review characteristics, so the number of features identified for the factor are far greater.

This work was supported by The Coding School's High School AI Research Program.

III. Data & Methodology

A. Dataset

1) Description: The dataset used is the IMDB movie reviews [5], containing columns for Review, Rating, and Sentiment. Of the initial 50,000 rows, 49,586 remained after removing duplicates and non-applicable rows. The dataset was originally prepared for sentiment analysis, so the dataset has intentionally had its neutral ratings and reviews removed.

2) Preprocessing & Feature Engineering: I engineered 18 of the 22 features outlined by Celli [3], dropping those not present in my data:

1) Punctuation (ap)
2) Commas (cm)
3) Exclamation marks (em)
4) 1st person singular pronouns (im)
5) Negative particles (ne)
6) Numbers (nb)
7) Parentheses (pa)
8) Prepositions (pp)
9) Pronouns (pr)
10) Question marks (qm)
11) Long words (words with length ¿6) (sl)
12) Self reference (1st person pronouns, singular and plural) (sr)
13) Swears (sw)
14) Type/token ratio (tt)
15) Words (wc)
16) 1st person plural pronouns (we)
17) 2nd person singular pronouns (yu)
18) Mean word frequency (mf)

For clarity, excluding tt and mf, the counts of each item on the list above were used.

Using a feature correlation heatmap (Fig. 1), I identified highly correlated features and applied feature agglomeration to produce a final set of 8 core features.

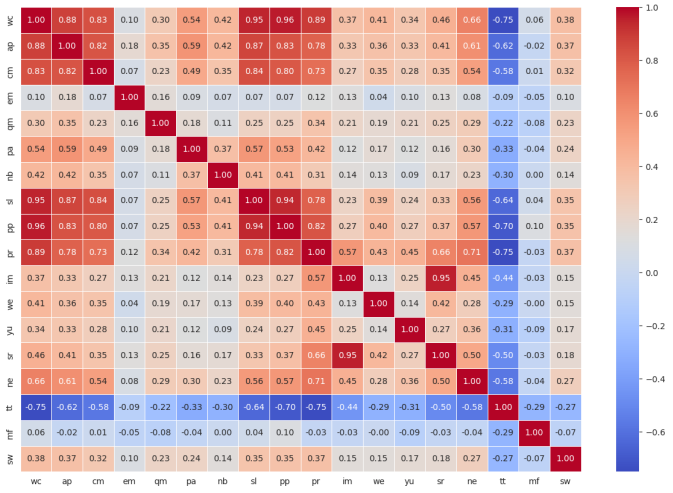


Fig. 1: Feature Correlation Heatmap.

B. Methods

I trained three unsupervised clustering models: K-Means (as a baseline), BIRCH (for scalability) [6], and GMM (to handle non-spherical clusters).

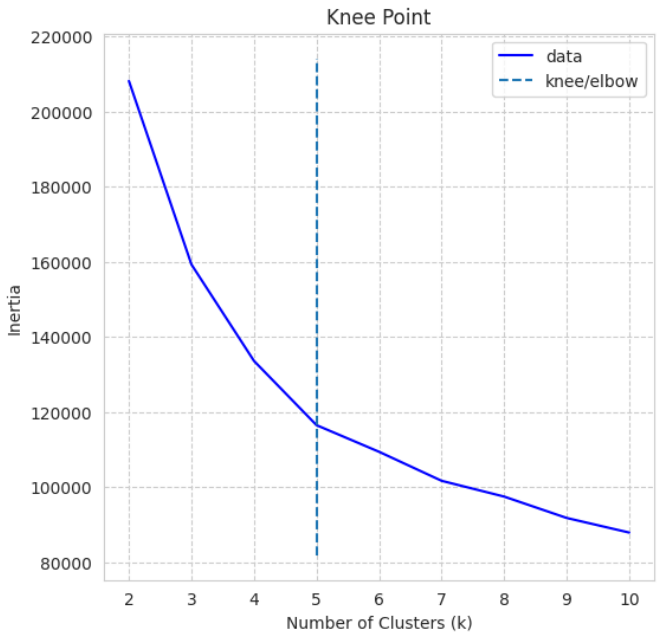


Fig. 2: Elbow Method for Optimal k.

For hyperparameter tuning, I used the elbow method on a K-Means model's inertia curve to find the optimal number of clusters ('k') (Fig. 2). This 'k' represents the number of personas and was used to train all three models. I then manually labeled each cluster by sampling its reviews. For evaluation, I used the silhouette score, Calinski-Harabasz score, and Davies-Bouldin index.

IV. Analysis & Findings

A. Model Evaluation

To evaluate model performance, a combination of standard clustering metrics was used. The results are summarized in Fig. 3, offering a comparison of how well each model separated the personas.

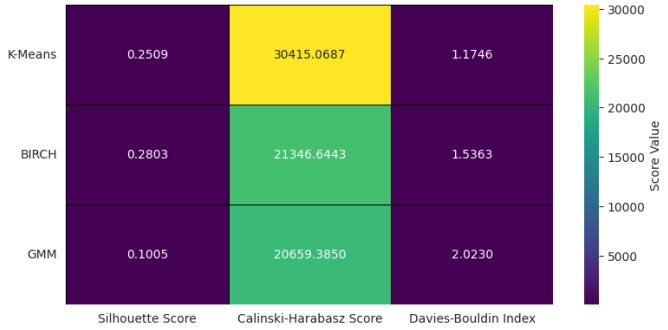


Fig. 3: Clustering Model Evaluation Scores.

B. Linguistic Personas

Based on the clusters, I identified 5 distinct reviewer personas present in the dataset, each with unique traits. The table of personas (Table I) details these findings. Interestingly, only the K-Means model was able to differentiate the emotional and candid personas; all other models merged the two personas to find only 4 personas. Below depicts further examinations.

1) Radar Charts (Fig. 4): The radar charts in Fig. 4 provide a visual comparison of the linguistic profiles for each persona as identified by the different models. There are a few new details:

- The academic dissector is almost a smaller version of the flaw ranter.
- The candid reviewers have a higher mean word frequency than others.
- The emotional persona tends to have a more unique vocabulary compared to other personas but lacks in every other factor.
- The structured critic seems to have a very balanced footprint.

2) UMAP Projection (Fig. 5): When graphing all 3 models using UMAP for dimensionality reduction and mapping the clusters to it, it is possible to see a lot of overlap between the personas [7]. However, all 3 models seem to be able to identify the same personas in generally similar locations. Interestingly, excluding the K-Means model, none of the other models were able to segment candid reviewers. Only the K-Means model was able to differentiate candid reviews, all other models combined it with emotional.

3) Rating Distribution (Fig. 6): Violin plots were used to plot the movie rating distribution across the personas, and they revealed interesting patterns. The strongest of which is that the emotional persona appears to give a very polarized rating. They avoid ratings closer to the median and favor the extremes.

4) Sentiment Distribution (Fig. 7): When graphing the average sentiment across personas, it is highly balanced.

5) Genre Distribution (Fig. 8): Personas seem to have a connection with the genre of movie being reviewed. Academic dissectors, candid reviewers, and structured critiques seem to review horror movies very often. Flaw ranters tend to review criminal based movies, and emotional reviewers commonly review comedy and romance movies.

V. Discussion

This research successfully congregates reviews based on their writing styles and analyzes each group (linguistic persona) created. This success has led to several discoveries, all of which have yielded potentially valuable interpretations and insights.

A key finding is the differential performance of the clustering models. Only the K-Means model successfully segmented the Candid persona from the Emotional one. This suggests that the geometric assumptions of K-Means, which partitions data into Voronoi cells, were better suited to separate the subtle stylistic differences between direct, blunt feedback (Candid) and purely emotive expression (Emotional). The other models, BIRCH and GMM, likely merged these personas due to their different approaches to defining cluster boundaries and density [6].

When examining the radar charts, the linguistic profiles reveal further nuances. The Academic Dissector appears as a less extreme version of the Flaw Ranter, suggesting both personas share an analytical approach, but the Flaw Ranter employs a more intense and critical tone. Another detail is that the Emotional persona shows a greater type/token ratio, indicating a more diverse vocabulary, which may be used to express a wide range of feelings. The Candid persona has the highest mean word frequency, hinting at the use of more common, straightforward language, which aligns with their direct and blunt style.

The rating and genre distributions provide behavioral context to these personas. The highly polarized ratings from the Emotional persona (Fig. 6) align with their emotionally driven reviews and preference for genres like comedy and romance, which often elicit strong positive or negative feelings. The Flaw Ranter's focus on crime movies is also logical, as these films often have intricate plots ripe for dissection and criticism. The overlap in horror movie reviews among Academic Dissectors, Candid reviewers, and Structured Critics suggests the genre invites varied forms of analysis, from technical critiques to straightforward reactions.

VI. Conclusion & Future Work

This study successfully demonstrates that unsupervised clustering based on stylometric features can effectively identify distinct linguistic personas within movie reviews. By moving beyond simple sentiment analysis, this research provides a more nuanced framework for understanding the diverse ways audiences express their opinions. The identification of five unique personas, Academic Dissector, Candid, Emotional, Flaw Ranter, and Structured Critic, each with characteristic linguistic patterns and behavioral tendencies, confirms that writing style is a rich source of information.

For future work, this research can be expanded in several promising directions. First, the feature set could be enriched by identifying more sophisticated linguistic cues that might have a greater impact on persona differentiation [2], [3]. Second, the model's adaptability should be tested in other domains, such as product or travel reviews, to see if similar or different personas emerge.

The practical applications of this work are extensive and varied:

- **Become a Smarter Shopper:** Users could filter reviews to focus on personas most helpful to them, such as prioritizing a "Structured Critic" over a "Flaw Ranter" to get a balanced view.
- **Improve Your Communication:** Individuals can gain insight into their own writing style, helping them to tailor their communication to be more effective and constructive.
- **Spot Online Bias:** This framework can help users see past purely emotional language and better evaluate the quality of an argument, distinguishing an "Emotional" outburst from a "Structured" point.
- **Navigate Online Spaces:** Understanding the dominant personas in a forum or social media group can help users quickly grasp the "vibe" of the community.
- **A Tool for Creativity:** Writers and creators could use these persona profiles to build more believable and linguistically consistent characters, ensuring, for example, that a detective character consistently speaks like a "Structured Critic."

Ultimately, this line of research can lead to more sophisticated tools for navigating and interpreting the vast landscape of online user-generated content.

References

[1] J. W. Pennebaker, "The secret life of pronouns," *New Scientist*, vol. 211, no. 2828, pp. 42–45, 2011.

[2] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore, "Using linguistic cues for the automatic recognition of personality in conversation and text," *Journal of artificial intelligence research*, vol. 30, pp. 457–500, 2007.

[3] F. Celli *et al.*, "Unsupervised personality recognition for social network sites," in *Proc. of sixth international conference on digital society*, 2012, pp. 59–62.

[4] H. Sharma and A. G. Aggarwal, "Segmenting reviewers based on reviewer and review characteristics," *International Journal of Business Analytics (IJBAN)*, vol. 9, no. 1, pp. 1–20, 2022.

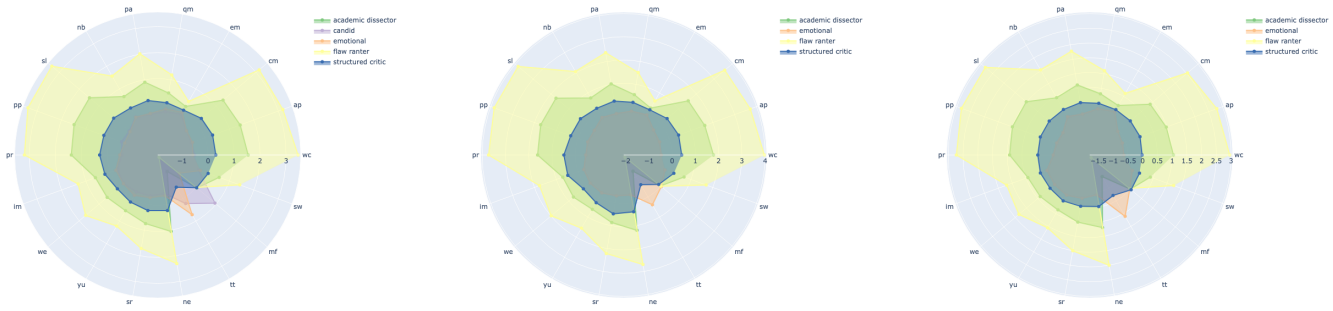
[5] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, June 2011, pp. 142–150. [Online]. Available: http://www.aclweb.org/anthology/P11-1015

[6] T. Zhang, R. Ramakrishnan, and M. Livny, "Birch: an efficient data clustering method for very large databases," in *Proceedings of the 1996 ACM SIGMOD international conference on Management of data*, 1996, pp. 103–114.

[7] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," 2018.

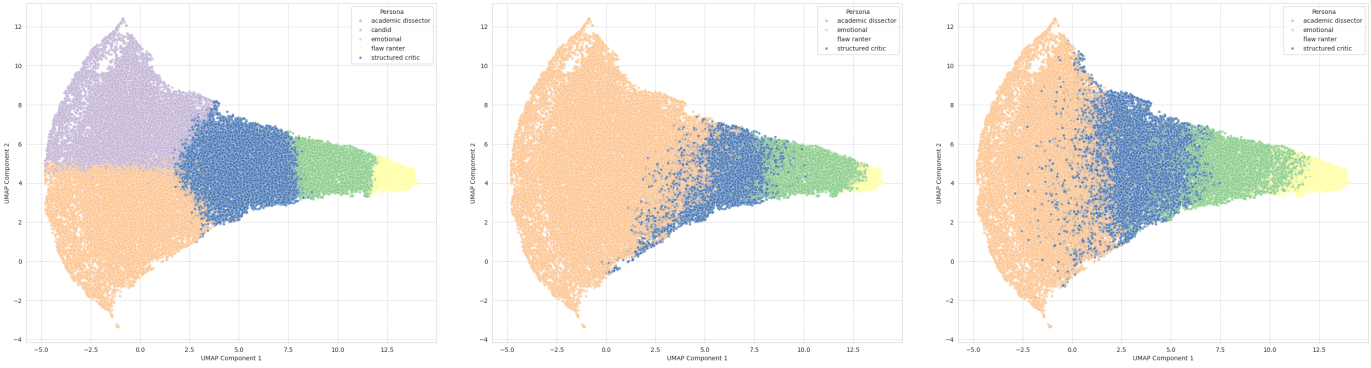
TABLE I: Description of Identified Linguistic Personas

Linguistic Persona	Description	Key Traits	How They Are Helpful
Academic Dissector	Reviewers who delve into the details of a film, often analyzing its historical context, technical aspects, or deeper thematic elements.	Detailed analysis, focus on context (historical, technical), objective tone, exploration of cinematic craftsmanship.	Prioritize their detailed critiques for deep dives into specific functionalities, technical issues, or missing features. Their input is excellent for product roadmap planning and iterative improvements.
Candid	Reviewers who offer blunt, direct assessments, mixing strong negative critiques with enthusiastic positive narrations, often with personal context.	Direct language, focus on specific issues (e.g., "obscene language"), personal context/preferences, immediate sense of enjoyment or disapproval.	Gives insight into straightforward reactions, highlighting specific aspects (both good and bad) that resonate directly with the reviewer's preferences or moral compass.
Emotional	Reviewers who express strong emotional responses, both positive and negative, often with subjective reasoning and clear recommendations.	Strong emotional language, personal connection to the film's impact, subjective reasoning, clear recommendations ("See it" or "Forget it").	Useful for understanding immediate, visceral reactions to a film and gauging overall audience sentiment. Their direct calls to action can guide quick viewing decisions.
Flaw Ranter	Reviewers who provide highly detailed, often frustrated, dissections of films, pointing out specific failures and what the film should have done.	Extensive plot summaries, detailed examples to support criticism, focus on predictability and clichés, use of vivid/sarcastic language to express disappointment.	Valuable for understanding the exact points of failure in a film, identifying common plot pitfalls, and seeing detailed deconstructions of narrative or character issues.
Structured Critic	Reviewers who offer analytical and structured criticism, breaking down their opinions into specific categories and often drawing comparisons.	Categorized analysis, comparative reasoning, justification of opinions with specific details, willingness to challenge common perceptions.	Provides a more organized and detailed understanding of a film's strengths and weaknesses across different aspects (e.g., direction, acting, script).



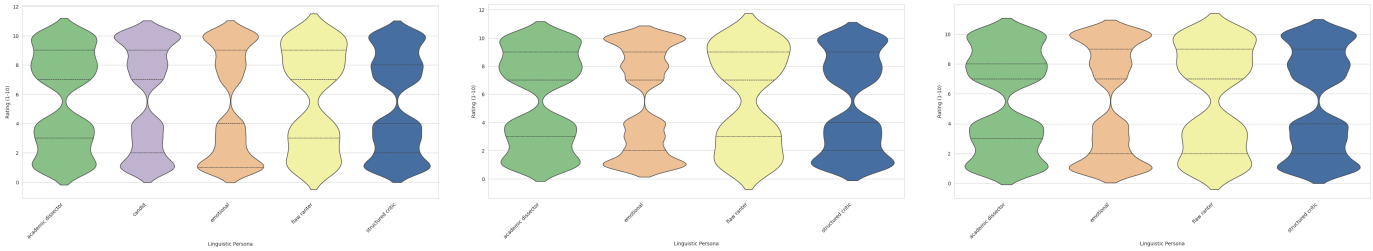
(a) K-Means Personas (b) BIRCH Personas (c) GMM Personas

Fig. 4: Linguistic Profile Radar Charts for Personas Across Models.



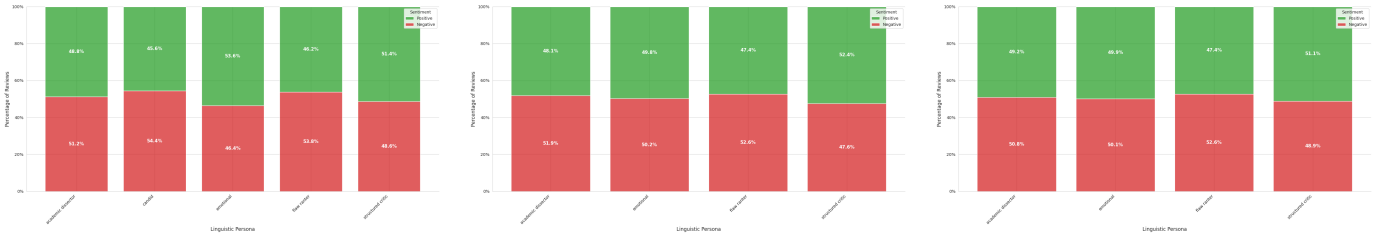
(a) K-Means Personas (b) BIRCH Personas (c) GMM Personas

Fig. 5: 2D UMAP Visualization of Personas by Clustering Model.



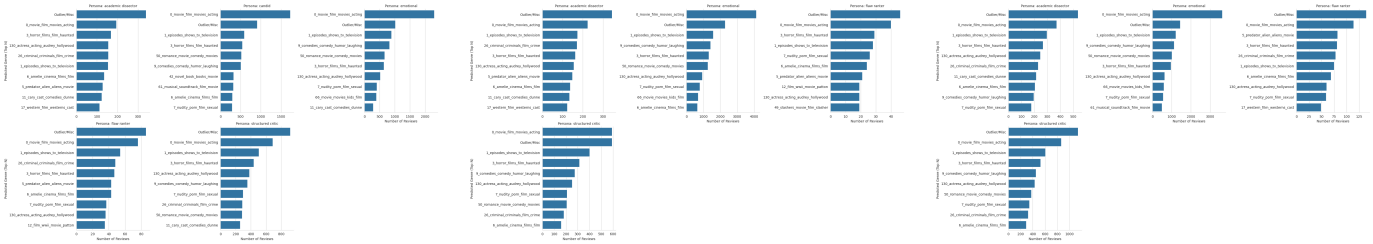
(a) K-Means (b) BIRCH (c) GMM

Fig. 6: Movie Rating Distribution by Persona for Each Model.



(a) K-Means (b) BIRCH (c) GMM

Fig. 7: Sentiment Distribution by Persona for Each Model.



(a) K-Means (b) BIRCH (c) GMM

Fig. 8: Top 10 Reviewed Genres by Persona for Each Model.