# Personality Analysis using PEML dataset

M. Abaeiani (st.nr: 2410235) , M. H. Mohammadirad (st.nr: 2409311) and K. Lund (st.nr: 2410693)

[1]University of Oulu, Finland
https://github.com/KristofferLund94/NLP_Personality_Analysis

**Abstract**

*This study explores the possibility of extracting personality traits through a conversational dataset using natural language processing. The personality traits we are exploring is the big five O.C.E.A.N traits. We examined how specific personality traits correlates with linguistic choices and conversational patterns. We make us of metrics such as vocabulary diversity, word frequency and word category analysis. Additionally machine learning models were implemented for the use of predicting personality traits, and then abstracting those traits to predict the speakers in the dataset.*

## 1. Introduction

This project investigates the intersection of emotion and personality in conversation, aiming to explore how specific personality traits correlate with emotional tendencies in dialogue exchanges. By analyzing the ways different personalities handle affirmation, repetition, and negation in their language, we can better understand how personality influences conversational patterns.

Our dataset, the Personality EmotionLine Dataset (PEML), contains paired conversations where each dialogue consists of three main utterances: Utterance_1 by Speaker_1, a response (Utterance_2) by Speaker_2, and a follow-up (Utterance_3) by Speaker_1. Each utterance is labeled with an emotion (such as neutral, surprise, fear, sadness, joy, disgust, or anger) and a sentiment (neutral, positive, or negative), providing insight into the emotional flow of the dialogue. Additionally, the personality traits of Speaker_1 are recorded for each exchange as an array of five values. These correspond to Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism.

Our objectives are twofold: first, to compile detailed profiles of each speaker by analyzing their vocabulary use, token frequencies, and specific word categories (such as confirmation and negation words). This gives us a numerical and linguistic view of each speaker's conversational style. Secondly, we aim to identify potential correlations between these conversational metrics and personality traits by observing how various personality dimensions affect language choices and emotional responses. For example, do speakers with higher levels of Extraversion use more affirming language? Do speakers with high Neuroticism tend to use more negations?

The outcome of this study could help develop NLP systems capable of adapting dialogue styles based on the inferred personality traits of a user, thereby enhancing conversational authenticity and user engagement.

## 2. Methodology

In this section, we present the methods employed to analyze the dataset and achieve the project's objectives. The project is divided into six distinct implementation tasks, each focusing on a specific aspect of the analysis. We will explore each of these tasks in detail, discussing the techniques, tools, and algorithms used, along with the rationale behind our methodological choices. This structured approach ensures that each task builds on the previous ones, collectively contributing to a comprehensive understanding of the dataset and fulfilling the project's overall goals.

### 2.1. General Trends

To explore the relationship between emotion, personality, and conversational patterns, we structured the analysis by first organizing data for each speaker. For each speaker, we gathered all utterances into a single file to analyze their vocabulary size, token usage, and certain recurring linguistic features. Each metric described below was calculated with specific focus on identifying any observable patterns that could relate to personality traits.

We implemented the following metrics for each speaker:

1. Vocabulary Size: We measured the variety of vocabulary each speaker used by counting the unique words across their utterances. This metric helped us gauge lexical diversity, which could reflect openness or creativity.

2. Total Tokens: This is the total word count used by each speaker across all their utterances. Total tokens give insight into the verbosity and communicative style of each speaker.

3. Repetitions: For this metric, we counted how frequently words were repeated within the same utterance. Higher repetition could indicate emphasis or particular conversational styles and may reflect aspects of extraversion or neuroticism.

4. Confirmations: We counted the occurrences of confirmation words, which often signal agreement or affirmation in dialogue. The confirmation words we considered were: “definitely,” “sure,” “yes,” “okay,” “ok,” “of course,” “certainly,” “indeed,” “yeah,” “absolutely,” “alright,” and “yep.” The frequency of these words can reveal a speaker’s level of agreeableness or extraversion.

5. Negations: We analyzed the use of negation words to identify any tendencies toward disagreement or avoidance. The negation words considered included: “no,” “not,” “nobody,” “never,” “isn’t,” “aren’t,” “don’t,” “doesn’t,” “wasn’t,” “weren’t,” “haven’t,” “hasn’t,” “won’t,” “hadn’t,” “neither,” “nor,” and “none.” A higher frequency of negations might signal a more neurotic or less agreeable personality style.

Additionally, to track how these conversational patterns evolved over the dialogue, we divided each speaker’s conversation into 500-token segments, or "buckets." Within each bucket, we calculated the count of repetitions, confirmations, and negations. This bucketing approach allowed us to observe whether certain tendencies intensified or diminished over time, potentially offering insights into how aspects of personality and emotion interact as the conversation progresses.

To gain a clearer understanding of each speaker’s personality profile, we averaged the scores provided for each of the five personality traits: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. By comparing these averages to our conversational metrics, we explored whether particular language patterns aligned with specific personality characteristics.

## 2.2. Topic Shifts

The goal of this analysis is to evaluate the extent to which a speaker changes topics during a conversation. We hypothesize that frequent changes in topic may be correlated with quick introductions of new vocabulary and a higher number of distinct topics. The analysis is conducted in two parts.

### 2.2.1. Vocabulary Growth with Respect to Tokens

In this part of the analysis, we aim to assess how a speaker’s vocabulary evolves over the course of a conversation by analyzing the cumulative vocabulary size in relation to the number of tokens they use. This can provide insight into how quickly a speaker introduces new words and thus shifts topics, as an increase in vocabulary size may signal the introduction of new topics or ideas.

The method iterates through each utterance of the speaker, adding the number of tokens in each utterance to a running total (‘cumulative_tokens‘). This total represents the cumulative number of tokens the speaker has used up until that point in the conversation.

As the tokens from each utterance are processed, the unique tokens encountered are added to a cumulative vocabulary set. The set data structure is used because it automatically eliminates duplicates, ensuring that only unique vocabulary is counted.

After each utterance, the vocabulary size is updated by checking the number of unique words in the vocabulary set (‘cumulative_vocabulary‘).

The cumulative number of tokens and the corresponding vocabulary size are stored in a dictionary (‘vocabulary_growth‘) with two keys: "Tokens" and "Vocabulary".

If a speaker shows rapid vocabulary growth, this might be indicative of a personality trait that enjoys exploring new topics or engaging in more complex discourse (Openness). Conversely, a slower vocabulary growth might reflect a speaker who prefers sticking to familiar topics.

### 2.2.2. Topic Modeling Using LDA

In this part, the objective is to apply Latent Dirichlet Allocation (LDA) [Ble03] topic modeling to analyze the conversational topics of each speaker. The goal is to determine the optimal number of topics for each speaker and assess the coherence of the topics. The approach involves several steps to model topics, evaluate them, and visualize the results.

For each speaker, the input data consists of a list of tokenized text from their utterances. Each utterance is represented as an array of tokens, which will be used as input for the topic modeling process. Before applying LDA, the tokenized texts for each speaker are processed using Gensim’s Dictionary and corpus structures. The dictionary maps each unique word to an integer ID, and the corpus is a bag-of-words representation of the text, where each document (considered as all utterances of each speaker) is converted into a list of word IDs and frequencies.

LDA is then applied using Gensim’s ‘LdaModel‘ function. LDA is an unsupervised machine learning algorithm that assumes each document is a mixture of several topics. The algorithm finds the underlying topics by analyzing the distribution of words across documents. The number of topics is an input parameter, which will be varied to find the optimal number of topics.

To evaluate the quality of the topics generated, the coherence score is calculated using Gensim’s ‘CoherenceModel‘. This score measures the interpretability of the topics. Higher coherence scores indicate that the words within a topic frequently appear together in the corpus, making the topics more meaningful.

To determine the optimal number of topics for each speaker, the script runs LDA with varying numbers of topics (ranging from 2 to 10). For each number of topics, the corresponding coherence score is calculated and stored. Once the optimal number of topics is determined, the topics are extracted from the LDA model. Each topic consists of a distribution of words, and the top words in each topic provide insight into the subject matter or themes of that topic.

The optimal number of topics is chosen based on the coherence score. A higher coherence score indicates better topic quality, meaning the topics are more distinct and meaningful in terms of the words they contain. By comparing the optimal number of topics and the actual topics themselves across different speakers, we can assess whether certain speakers tend to cover more topics or stick to fewer, more focused themes.

The number of topics and the coherence of these topics may be linked to personality traits such as openness, extraversion, and agreeableness. For example, extraverts may show more topics, indicative of a broader range of interests and more diverse conversations. Introverts might display fewer topics, suggesting more focused and less varied conversations. A speaker with high openness might introduce a greater variety of topics or show greater diversity in the words they use.

### 2.3. Emotion Pattern and Personality

This analysis aims to explore the relationship between emotional patterns and personality traits using the NRC Emotion Lexicon https://github.com/Priya22/EmotionDynamics/tree/master/lexicons. Specifically, we aim to determine whether the valence, arousal, and dominance (VAD) values can provide insights into the personality traits of different speakers. For this purpose, we calculate the average, standard deviation, minimum, and maximum values for each of these emotional dimensions based on the lexicon terms present in each speaker's conversation. Below is a breakdown of the approach and findings.

We use the NRC Emotion Lexicon which assigns values for Valence, Arousal, and Dominance to approximately 20,000 words. These values represent:

- Valence: The positive or negative emotional tone of a word.

- Arousal: The intensity of the emotion expressed by a word.

- Dominance: The level of control or dominance the word implies in the emotional context.

The script extracts and processes the tokens from the transcripts of each speaker in the conversation. For each speaker, it identifies which words are present in the NRC lexicon and calculates the average, standard deviation, minimum, and maximum for each of the VAD dimensions.

### 2.4. Dominant Personality Analysis

In this section, we focus on segmenting the dataset according to each speaker's dominant personality trait, which is indicated in the dataset labels. Each personality trait corresponds to one of the five factors in the OCEAN model: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. For each individual, the dominant trait is represented by a specific index, allowing us to group speakers based on the most prominent aspect of their personality. [McC99]

After segmenting the dataset by dominant personality traits, we perform a series of linguistic analyses to explore variations in speech patterns across different personality groups. This analysis centers on five key linguistic features: vocabulary size, which measures the diversity of words used; total token count, indicating overall word count; repetition frequency, capturing the rate of repeated words or phrases; confirmation frequency, reflecting instances of agreement or affirmation; and negation frequency, indicating occurrences of disagreement or denial.

To analyze these features, we use a bucket-based approach. This method involves dividing the dataset into smaller, sequential segments (buckets) and moving each bucket incrementally through the dataset. For each bucket, we calculate the occurrence rates of confirmation words, negation words, and repetitions, allowing us to track changes in these linguistic features over time within each personality trait segment. The algorithms we employ to calculate these metrics are similar to those described in Section 1 (General Trends), ensuring consistency in our approach.

By observing patterns across the buckets, we gain insights into the evolution and frequency of specific speech behaviors in relation to each dominant personality trait. The results obtained in this section reveal potential correlations between personality traits and speech tendencies. For instance, individuals with Openness as their dominant trait may display distinct patterns in the frequency of confirmation words compared to those whose dominant trait is Extraversion. This type of analysis helps highlight how personality influences language use, providing deeper insights into how people with different personality profiles express themselves.

### 2.5. Embedding-Based Personality Analysis

In this section, we perform a personality analysis based on embedding similarity. Our approach uses a doc2vec model to generate vector embeddings for each speaker's sentences, focusing specifically on the "utterance 1" section of the dataset to maintain consistency. To create a comprehensive document representation for each speaker, we combine all of their sentences into a single unified text document. The doc2vec model then infers a unique vector embedding for each speaker based on this compiled document, providing a dense numerical representation of their speech patterns.

To define personality traits, we rely on pre-established personality dictionaries aligned with the OCEAN model, encompassing Openness, Conscientiousness, Extraversion, Agreeableness, and Emotional Stability (the inverse of Neuroticism). Each trait is represented by multiple pairs of tuples, with each tuple containing one term for the positive side of the spectrum and another for the negative side. For example, the Extraversion trait includes pairs such as ("Talkative," "Silent"), ("Sociable," "Reclusive"), and ("Adventurous," "Cautious").

For each trait, we use these tuples to derive multiple positive and negative embedding vectors. These vectors collectively represent each personality trait from both ends of the spectrum, allowing us to capture subtle distinctions in language that might be missed if only one side were considered.

With these embeddings in place, we calculate the cosine similarity between each speaker's vector and the vectors representing each personality trait's positive and negative sides. This similarity score provides a measure for each OCEAN trait, indicating the extent to which the speaker aligns with either the positive or negative end of the personality spectrum.

To evaluate the model, we compare the personality scores derived from our embeddings with the original personality trait labels. This allows us to assess the model's accuracy in predicting personality traits based on language use.

To further test the robustness of our approach, we conduct experiments using two variations of the personality dictionaries: one with an extended vocabulary for each trait and another that excludes terms associated with the negative side of the spectrum. By comparing the results across these different configurations, we can better understand how vocabulary choices impact the model's accuracy and interpretability.

### 2.6. Machine Learning Model

We want to explore the use of machine learning to predict the personality traits for 'Speaker_1' per row in the dataset. Our goal is to explore how preprocessing the dataset is affecting the machine learning model. There is abundant existing research about this already, but for this task we are doing our own. [MPGC17] [PBCP20]

The first step is to do tf-idf vectorization on the utterances, where we do not combine the three utterances beforehand, as we want to keep the information of who is speaking and in what order.

Predicting five personality scores is a regression problem, therefore we have the choice of using different regression models; polynomial regression, support vector regression, decision tree, random forest, deep learning, to name a few. We started with the goal of comparing all these models, however, doing grid search on all these models proved to use too much time for the time we had to spend on this project. Since the project is a natural language processing project, our time is better spent on optimizing the preprocessing rather than the model.

Therefore, we are focusing on a sequential deep learning model, where we do a simple grid search for the optimal parameters. We split the dataset into training, validation and testing, and measure the performance with mean squared error. The best parameters we found were a batch size of 32, 10 epochs, 256 dense nodes and a dropout rate of 05.

As our main objective is to explore how different alterations of the dataset affect the performance of the model, we want to make small modifications of the dataset and measure it up against the original data. We explore limiting the vocabulary of the model to 5000 tokens and 1000 tokens. We try with some N-grams; bigram, trigram and quadgram. We also try removing the emotion labels, sentiment labels, and both. We investigate keeping only utterance 1 and keeping only it's emotion and sentiment label. Lastly, we omit only utterance 2, since this is speaker_2, alongside it's labels.

After getting the results, we combine the successful features and test datasets further based on the results.

### 2.7. Abstracting Predicted Personality Scores to Speakers

The project does not ask for it, but we thought it would be interesting to see if we could take the five predicted personality traits and measure the average euclidean distance to each speaker's average personality scores to find the closest speaker to the prediction. The results of this is shown further down in the document.

## 3. Rresults

### 3.1. Task 1: General Trends

The table below shows the vocabulary size, total tokens, repetitions, confirmations, and negations for each speaker:

Each speaker's language use shows interesting variations in vocabulary size, total tokens, word repetitions, confirmations, and negations:

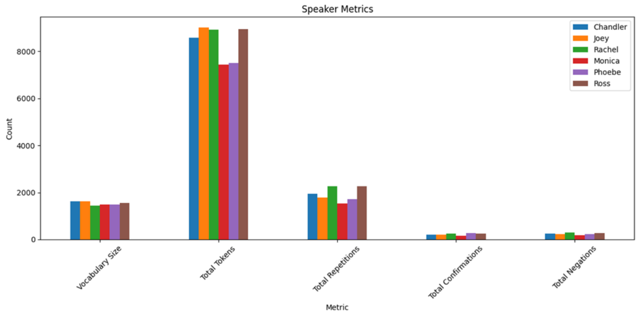


**Figure 1:** *Speaker metrics*

- Vocabulary Size: All speakers display a broad vocabulary, with sizes ranging from 1429 to 1621 unique words. Joey and Chandler lead with 1621 and 1620 words, respectively, while Rachel has a comparatively smaller vocabulary at 1429 words. These figures indicate that Joey and Chandler's dialogues are slightly richer in lexical variety.

- Total Tokens: Token usage varies, with Joey using the most tokens (9018) and Monica using the least (7448). This shows that Joey tends to have longer or more frequent utterances compared to other characters, potentially reflecting his more conversational style.

- Repetitions: Repetitions per speaker reveal an interesting trend. Ross and Rachel lead with 2272 and 2260 repetitions, respectively, while Monica has the fewest at 1528. This high count for Ross and Rachel suggests a tendency for these characters to emphasize their points or revisit phrases, potentially indicative of certain conversational habits or emotional states.

- Confirmations: The number of confirmation words are highest in Rachel and Phoebe's dialogues (259 and 261, respectively), hinting at their conversational openness or a more affirming communication style. Monica, on the other hand, uses the fewest confirmation words (155), indicating a possible preference for more direct or assertive expressions.

- Negations: Negation usage is relatively balanced, with Rachel and Ross using them most frequently (292 and 279, respectively), and Monica the least (188). This suggests Rachel and Ross may have a more cautious or critical tone in conversation.

#### 3.1.1. Personality Insights

To analyze how personality traits might relate to conversational behavior for each speaker, we should examine personality averages:

- Openness: Ross and Monica exhibit high openness scores (0.722 and 0.713, respectively), which aligns with their tendency toward varied vocabulary, as people with high openness often engage in more exploratory or idea-rich dialogues.

- Conscientiousness: Joey scores notably high in conscientiousness (0.614), while Rachel has one of the lower scores (0.354). This may suggest Joey's more structured conversational style and might explain his frequent token use and vocabulary size, while Rachel's

**Table 1:** *Speaker Metrics Summary*

| Speaker | Vocabulary Size | Total Tokens | Total Repetitions | Total Confirmations | Total Negations |
|---|---|---|---|---|---|
| Chandler | 1620 | 8587 | 1952 | 204 | 250 |
| Joey | 1621 | 9018 | 1790 | 204 | 237 |
| Rachel | 1429 | 8923 | 2260 | 259 | 292 |
| Monica | 1492 | 7448 | 1528 | 155 | 188 |
| Phoebe | 1479 | 7511 | 1726 | 261 | 233 |
| Ross | 1557 | 8961 | 2272 | 254 | 279 |

**Table 2:** *Personality Traits Average Scores for Speakers*

| Speaker | Openness | Conscientiousness | Extraversion | Agreeableness | Neuroticism |
|---|---|---|---|---|---|
| Chandler | 0.648 | 0.375 | 0.386 | 0.580 | 0.477 |
| Joey | 0.574 | 0.614 | 0.297 | 0.545 | 0.455 |
| Rachel | 0.635 | 0.354 | 0.521 | 0.552 | 0.469 |
| Monica | 0.713 | 0.457 | 0.457 | 0.660 | 0.511 |
| Phoebe | 0.600 | 0.480 | 0.310 | 0.460 | 0.560 |
| Ross | 0.722 | 0.489 | 0.600 | 0.533 | 0.356 |

less structured approach could be linked to her higher confirmation and negation rates.

- Extraversion: Ross shows the highest extraversion score (0.600), which aligns with his high token count and repetition rate, as extroverts tend to communicate more openly and repetitively to engage their audience. Joey and Phoebe, who score lower in extraversion, use fewer tokens and may appear slightly more reserved or less talkative.

- Agreeableness: Monica's high agreeableness score (0.660) may correlate with her low use of confirmations and negations, suggesting a tendency for straightforward communication without frequent reassurances or hesitations.

- Neuroticism: Phoebe scores highest in neuroticism (0.560), which might relate to her higher-than-average use of confirmation words, potentially reflecting a need for affirmation or reassurance.

### 3.1.2. Bucketing

To analyze the evolution of repetition, confirmation, and negation use as conversations progress, we divided each speaker's total token count into buckets of 500 tokens. Each bucket groups a sequence of 500 tokens in the order they appeared in the conversation, allowing us to observe how certain linguistic features change with increased conversational length.

For each bucket, we calculated the number of repetitions, confirmations, and negations. This bucketing approach provides a clearer picture of linguistic dynamics, as it helps track whether certain behaviors—like repeating phrases or using affirmations—occur more frequently as conversations get longer. For instance, we might observe whether speakers increase confirmations or negations to manage the flow or tone of a longer exchange.

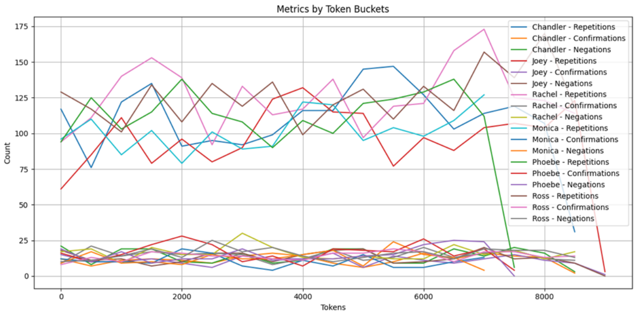


**Figure 2:** *Metrics by token buckets*

1. Repetitions: Most speakers show fluctuations in repetition use across buckets, with some characters like Rachel and Ross showing higher consistency in maintaining frequent repetitions throughout. Ross consistently uses a high number of repetitions, peaking in buckets like 7000 with 157 instances. Chandler also shows high repetition counts, particularly in later buckets (e.g., 5000 and 5500). Characters such as Joey and Phoebe have more variable repetition counts, with peaks in specific buckets (e.g., Joey in 3500 with 124 repetitions).

2. Confirmations: Confirmation usage varies significantly across speakers. Phoebe tends to increase confirmations in the middle and later buckets, with peaks around 2000 and 6000 tokens. This trend might suggest a greater need for assurance or affirmation as conversations lengthen. Rachel also shows a notable rise in confirmation usage, especially in buckets like 7000, where it reaches 16 instances, indicating an affirming conversational style that intensifies in longer exchanges. By contrast, Monica and Ross have relatively moderate confirmation usage, with only occasional peaks, such as Monica's bucket 2500 with 16 confirmations.

3. Negations: Negation usage remains fairly stable across most characters, although Rachel and Ross show occasional increases in later stages. For example, Rachel has a rise in negations around bucket 6500 (22 instances), which could imply a more critical or

oppositional tone as the conversation develops. Joey and Monica, on the other hand, exhibit more sporadic negation counts, with some peaks in specific buckets (e.g., Joey's 7000 bucket with 20 instances), possibly reflecting moments of emphasis or disagreement.

Overall, the bucketing analysis reveals that some characters, like Ross and Rachel, demonstrate consistent behaviors in repetitions and confirmations that align with their communicative styles. Phoebe and Joey show more variability, perhaps indicating a conversational approach that adapts more depending on the context or the stage of dialogue. These patterns underscore the different ways each character manages conversational flow, with certain personalities showing more stability and others adapting dynamically across extended interactions.

### 3.1.3. Task1 Conclusion

This analysis demonstrated a clear link between personality traits and linguistic tendencies in dialogue systems. Specifically, openness is associated with a larger vocabulary, neuroticism with higher negation rates, and conscientiousness with structured, lower-repetition dialogue. These findings highlight the potential of integrating personality traits into dialogue system design, leading to more personalized and engaging interactions.

## 3.2. Task 2: Topic Shifts

### 3.2.1. Vocabulary Growth with Respect to Tokens

The results reveal key differences in the token count and vocabulary diversity for each character's speech across utterances. The data indicates how many tokens and distinct vocabulary items each character uses over time. The findings provide insights into each character's language patterns and how these relate to personality traits.

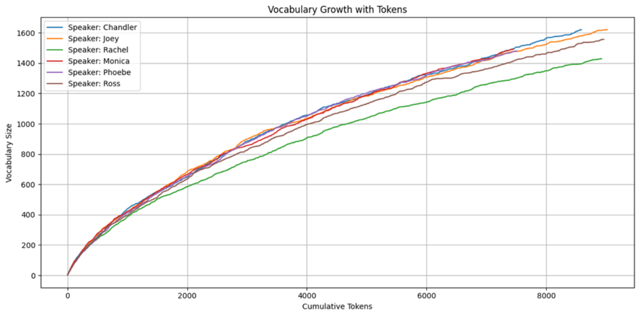


**Figure 3:** *Vocabulary growth with tokens*

As we can see, Chandler and Joey exhibit a gradual increase in token and vocabulary counts across episodes, reflecting high verbosity and linguistic diversity. Chandler peaks at 8587 tokens with 1620 unique words, and Joey reaches 9018 tokens and 1620 unique words, which suggests they use varied language throughout the episodes. This diversity and verbosity align with high Extraversion, as both characters often engage in lively conversations and display a broad range of expressions. Chandler's wit and Joey's expressiveness further highlight their open, socially engaging personas.

Phoebe and Monica show moderate token counts and vocabulary diversity, with Phoebe peaking at 7511 tokens and 1479 unique words, and Monica displaying similar values. Despite speaking less, both characters maintain rich vocabularies, which aligns with Agreeableness and Openness to Experience. Phoebe's unique and creative language reflects high Openness, while Monica's structured and emotionally supportive dialogue style suggests a balance between Openness and high Agreeableness.

Ross has a high token count, reaching 8961 tokens, but a comparatively low vocabulary diversity with 1557 unique words. This pattern suggests Conscientiousness in Ross, as he often speaks at length but with a constrained vocabulary, likely reflecting his preference for precision and consistency. Rachel displays a high number of tokens at 8923 but with only 1429 unique words, suggesting limited vocabulary diversity. This could be linked to Agreeableness and Extraversion. While Rachel is highly social and conversational, her repetitive language may reflect a tendency to focus on familiar themes or expressions, aligning with her agreeable and sometimes straightforward personality.

Overall, these observations align with character traits. Extroverted characters like Chandler and Joey demonstrate more linguistic diversity, while characters such as Ross and Rachel display more constrained vocabulary use, consistent with their personalities. By examining token and vocabulary counts, we gain insights into each character's language evolution across episodes, with implications for their personality traits and the thematic diversity of their dialogue.

### 3.2.2. Topic Modeling Using LDA

In Part 2, Latent Dirichlet Allocation (LDA) was applied to determine the optimal number of topics and assess topic coherence for each character in the data. This analysis aimed to reveal the diversity and thematic focus in each character's dialogue.

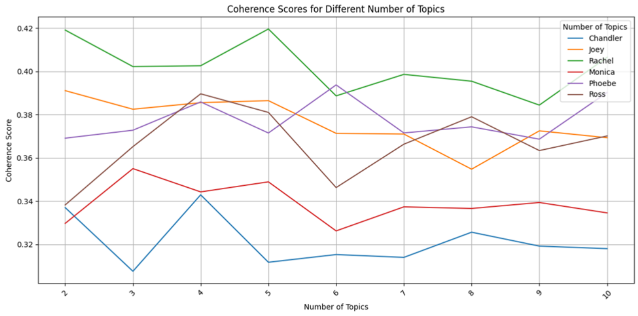


**Figure 4:** *Coherence scores for speakers*

The coherence scores for each number of topics across characters suggest varied trends. Chandler achieved his highest coherence score at 0.3429 with four topics, indicating that his dialogue was best represented with this number, beyond which coherence declined slightly. Monica's highest coherence of 0.3551 appeared with three topics, suggesting a moderate level of thematic diversity in her dialogue. Phoebe's dialogue showed a more complex distribution, achieving the highest coherence of 0.3938 with six topics, reflecting her eclectic conversational style. Ross displayed his peak coherence of 0.3896 with four topics, suggesting a middle ground between focus and variety. Joey's scores were generally stable across two to ten topics, peaking at two topics with a coherence of 0.3911, suggesting a narrower, more focused conversational range. Finally, Rachel's dialogue achieved its highest coherence of 0.4196 with five topics, indicating a more varied conversational scope.

Additionally, the best topics for each character provide insights into their conversational styles. Chandler's topics reveal a strong use of "I," "you," and "what," reflecting direct and interpersonal exchanges. Joey's topics are more straightforward, with common phrases like "hey," "right," and "know," aligning with his casual, humorous nature. Rachel's dialogue topics contain more emotional and reflective terms, such as "just," "oh," and "yeah," supporting her expressive and socially dynamic personality. Monica's topics are detail-oriented, with frequent use of "you," "what," and "it," reflecting her often instructive and problem-solving dialogue. Phoebe's topics are diverse, including unique terms like "really" and "okay," which align with her unconventional personality. Ross's topics include reflective words like "no," "okay," and "that," indicating a mix of intellectual and emotional themes in his dialogue.

**Chandler:**

- Topic 1 → i, you, that, oh, was, of, in, with, me, what
- Topic 2 → you, in, okay, have, like, oh, yeah, well, i, just
- Topic 3 → you, what, that, i, we, so, it, me, joey, are
- Topic 4 → i, it, you, no, hey, this, so, what, know, oh

**Joey:**

- Topic 1 → i, you, it, that, hey, know, what, out, me, right
- Topic 2 → you, i, it, yeah, oh, hey, that, no, what, this

**Rachel:**

- Topic 1 → i, you, no, yeah, hey, what, so, just, right, well
- Topic 2 → i, oh, that, you, this, it, just, so, not, but
- Topic 3 → i, you, oh, that, what, m, me, just, my, yeah
- Topic 4 → you, it, what, that, i, just, okay, no, of, me
- Topic 5 → you, i, me, oh, look, at, hi, what, my, this

**Monica:**

- Topic 1 → i, you, that, oh, it, are, me, what, out, my
- Topic 2 → you, what, it, this, okay, of, oh, yeah, so, hey
- Topic 3 → you, i, it, of, know, in, have, we, me, with

**Phoebe:**

- Topic 1 → i, you, we, but, that, just, so, really, this, okay
- Topic 2 → i, that, no, you, oh, my, it, in, here, this
- Topic 3 → you, no, so, okay, i, it, do, are, well, me
- Topic 4 → i, you, that, what, of, so, it, yeah, just, was
- Topic 5 → i, it, hey, know, no, you, oh, do, not, so
- Topic 6 → you, oh, yeah, i, what, are, my, okay, he, this

**Ross:**

- Topic 1 → you, i, no, what, that, it, oh, on, okay, me
- Topic 2 → i, you, it, yeah, this, on, she, that, have, can
- Topic 3 → i, you, we, it, no, m, have, that, do, okay
- Topic 4 → i, you, it, me, that, hey, was, just, of, well

Overall, these results indicate that the characters exhibit distinctive conversational patterns. Characters with more coherent topics at fewer numbers, like Joey, tend to have focused and consistent dialogue styles, while characters like Rachel and Phoebe with higher coherence scores at larger topic counts suggest greater thematic variety. This variation in topic number and coherence could potentially link to personality traits such as extraversion, openness, and agreeableness, where characters with more topics and higher coherence may demonstrate broader interests and more dynamic conversational styles.

### 3.2.3. Evaluation of Suitability for Discriminating Personality Traits

Vocabulary Growth Approach: This approach is useful for understanding the diversity of language employed during a conversation, but it may not directly reflect personality traits in a nuanced manner. For example, speakers with a broad range of vocabulary might appear more expressive, but this doesn't necessarily correlate with personality traits like openness or extroversion. However, in combination with other behavioral data, vocabulary growth could provide a rough proxy for engagement levels or a propensity for shifting topics.

LDA Topic Modeling Approach: LDA modeling is a more sophisticated method for evaluating topic shifts and can reveal deeper patterns in the conversation. By determining the optimal number of topics and examining the content of those topics, we can infer personality traits related to cognitive processes, interests, or conversational style. For instance, a speaker with a high number of topics and diverse content might be seen as more open-minded or socially engaged, whereas a speaker with fewer, more consistent topics might suggest a more focused or reserved personality. However, topic modeling is still limited by the granularity of the data and the interpretation of topic content. More personalized approaches might be needed to truly correlate topic diversity with personality traits like extraversion, agreeableness, or emotional expressiveness.

### 3.2.4. Task 2 Conclusion

Both methods provide interesting insights into the way speakers engage in conversations, with vocabulary growth indicating shifts in language use and topic modeling revealing underlying themes. These approaches have their merits in analyzing personality traits, but further refinement and integration of more behavioral data would be necessary to make more robust inferences about personality.

## 3.3. Emotion Pattern and Personality

Table below summarizes the VAD scores for each speaker, with a bar chart for visual comparison.

**Valence (Emotional Tone):** - Valence Average: Scores indicate that all speakers have relatively high emotional positivity, with values ranging from 0.571 (Chandler) to 0.598 (Joey). This suggests a generally positive emotional tone in their speech.

| Speaker | V_Avg | V_Std | V_Min | V_Max | A_Avg | A_Std | A_Min | A_Max | D_Avg | D_Std | D_Min | D_Max |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Chandler | 0.5706 | 0.2301 | 0.010 | 1.0 | 0.4609 | 0.1787 | 0.071 | 0.945 | 0.4918 | 0.1747 | 0.081 | 0.943 |
| Joey | 0.5978 | 0.2146 | 0.020 | 1.0 | 0.4551 | 0.1760 | 0.073 | 0.945 | 0.4990 | 0.1724 | 0.120 | 0.972 |
| Rachel | 0.5887 | 0.2328 | 0.020 | 1.0 | 0.4629 | 0.1868 | 0.082 | 0.952 | 0.4919 | 0.1752 | 0.113 | 0.981 |
| Monica | 0.5861 | 0.2248 | 0.010 | 1.0 | 0.4593 | 0.1874 | 0.046 | 0.951 | 0.4826 | 0.1721 | 0.081 | 0.981 |
| Phoebe | 0.5786 | 0.2287 | 0.010 | 1.0 | 0.4630 | 0.1840 | 0.090 | 0.971 | 0.4919 | 0.1684 | 0.087 | 0.963 |
| Ross | 0.5943 | 0.2273 | 0.005 | 1.0 | 0.4665 | 0.1861 | 0.073 | 0.959 | 0.4982 | 0.1755 | 0.081 | 0.963 |

**Table 3:** *VAD Scores for Each Speaker*

- Valence Standard Deviation: All speakers show relatively consistent emotional tone across their conversations, with slight variations in standard deviations. For example, Chandler's conversation is slightly more variable in emotional tone (0.230) compared to Joey (0.215).

- Valence Min/Max: The minimum valence values are fairly low across the board, but they do not show extreme negative emotions, with the minimum value being as low as 0.005 (Ross). This highlights that even in the most negative moments, their speech remains moderately positive.

**Arousal (Emotional Intensity):**

- Arousal Average: Scores are relatively uniform across all speakers, with values around 0.46. However, Phoebe's arousal score is slightly higher (0.463), indicating that her conversations might be more intense or energetic.

- Arousal Standard Deviation: Arousal varies across speakers, with Monica and Rachel showing slightly higher variability (0.187) compared to others, suggesting more fluctuating emotional intensity in their speech.

- Arousal Min/Max: The minimum and maximum arousal values range from 0.046 (Monica) to 0.090 (Phoebe), indicating some speakers experience less intense emotions in certain conversations, while others might show moments of high emotional intensity.

**Dominance (Control in Interaction):**

- Dominance Average: The average dominance scores suggest that all speakers exhibit relatively similar levels of control in their conversations, with values ranging from 0.492 (Chandler, Phoebe, Ross) to 0.499 (Joey). This implies that the speakers maintain a comparable sense of control and assertiveness.

- Dominance Standard Deviation: The dominance standard deviation is quite stable across speakers, with values ranging from 0.168 to 0.175. This indicates that the control in their conversations does not vary dramatically.

- Dominance Min/Max: The minimum and maximum values for dominance are similar across speakers, indicating a consistent pattern of dominance in their speech, with no extreme highs or lows.

### 3.3.1. Task 3 Conclusion

The analysis of VAD scores reveals several key insights into the emotional patterns of different speakers. However, the findings are relatively uniform across the different speakers, with only slight differences in arousal intensity and dominance.

- Valence seems to be quite consistent, showing a generally positive tone in the conversations, which may suggest a personality trait characterized by optimism or sociability.

- Arousal varies more significantly across speakers, with Phoebe showing the highest emotional intensity, which could be linked to a more expressive or spontaneous personality.

- Dominance is relatively consistent across speakers, suggesting that their interactions maintain a balance of control, with no speaker exhibiting extreme dominance or submission.

From these findings, we can infer that while VAD scores provide valuable insight into the emotional nature of conversations, they may not offer a strong discriminatory basis for identifying distinct personality traits. Further analysis, perhaps combining VAD with other linguistic features (e.g., sentiment, engagement), might be necessary to draw more definitive conclusions about the link between emotion patterns and personality.

### 3.4. Dominant Personality Analysis

The table below shows the result after dividing the dataset based on dominant personality trait.

| Dominant Personality | Count |
|---|---|
| 0 | 5387 |
| 1 | 1123 |

**Table 4:** *Counts for Dominant Personality*

As we can see, out of the five personality traits, only the first two traits are dominant among all speakers. The most prevalent trait, Openness, accounts for nearly 82.7% of the total data, while the second dominant trait, Conscientiousness, comprises the remaining 17.3%. This distribution suggests a potential bias in the analysis, as the data is unevenly distributed across traits, with no records for the other three traits. Keeping this limitation in mind, we proceed with the analysis of other metrics in the following sections.

### 3.4.1. Personality Insights

The table below shows vocabulary size, total tokens, repetitions, confirmations, and negations for each dominant personality, since the data was not evenly distributed, we also considered the ratio of negations to confirmations and the ratio of repetitions to total tokens for analysis.

| Idx | Vocab Size | Tokens | Reps | Confirms | Negs | Neg/Conf | Rep/Tok |
|---|---|---|---|---|---|---|---|
| 0 | 4022 | 41430 | 3234 | 1011 | 821 | 0.812 | 0.078 |
| 1 | 1621 | 9018 | 576 | 194 | 154 | 0.794 | 0.064 |

**Table 5:** *Vocabulary and Token Analysis with Ratios*

As we observe from the data, the ratio of negation to confirmation is slightly higher among posts with dominant openness. This trend may suggest that individuals with high openness are more inclined to express disagreement when encountering something they dislike. The trait of openness often correlates with a willingness to voice diverse opinions, which could explain the higher rate of negations in this group.

Additionally, the data reveals that the ratio of repetitions is notably higher in posts with dominant openness. This could indicate that those who are more open tend to emphasize or restate information more frequently when they feel it is necessary. This pattern aligns with the characteristics of openness, as such individuals may be more inclined to repeat or clarify ideas to ensure effective communication and understanding.

### 3.4.2. evolution analysis

Here are the graphs for evolution of negation, confirmation and repetitions for each dominant personality trait.

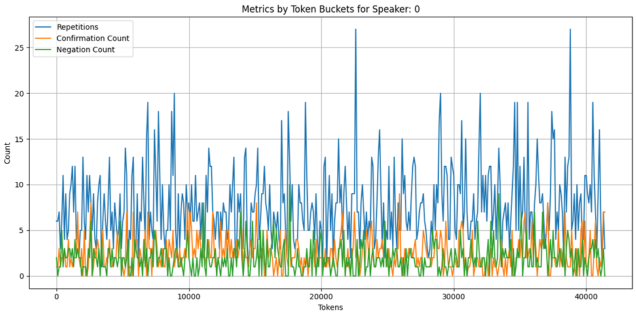


**Figure 5:** *Openness metrics evolution*

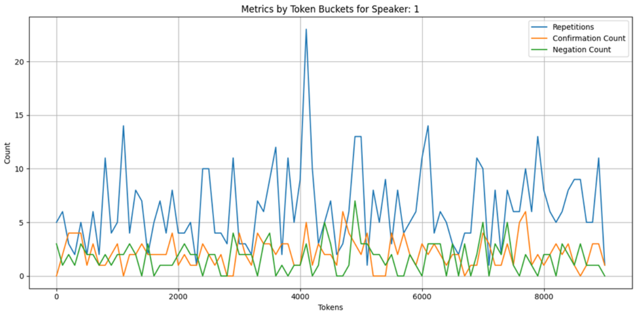


**Figure 6:** *Conscientiousness metrics evolution*

The graph shows that speakers with dominant openness consistently reach peaks of over 15 repetitions per 500 words in each bucket, reinforcing our earlier suggestion that openness is linked to a greater tendency for clarification through repetition. This pattern suggests that highly open individuals are more likely to emphasize or restate ideas to ensure clarity. Additionally, the graph indicates that both negations and confirmations peak more frequently among those with high openness, aligning with our previous observations. This finding highlights how openness may lead to more expressive communication, with individuals readily affirming or disagreeing with ideas to convey their perspectives effectively.

### 3.5. Embedding-Based Personality Analysis

In this section, we present the results of the analysis using embedding models and compare these results to the actual labels provided in the dataset. The aim is to evaluate how effectively the embeddings can capture the personality traits based on the provided indicators. Personality Dictionary for Embeddings For the embedding models, we used a basic personality dictionary, which contains key indicators for each trait. These indicators are essential for computing the similarity between documents (or speakers) and the traits. The following table shows the positive and negative indicators associated with each personality trait:

| Trait | Indicator (Positive / Negative) |
|---|---|
| **Extraversion** | Talkative / Silent<br>Sociable / Reclusive<br>Adventurous / Cautious<br>Open / Secretive |
| **Agreeableness** | Good-natured / Irritable<br>Cooperative / Negativistic<br>Mild / Headstrong<br>Not jealous / Jealous |
| **Conscientiousness** | Responsible / Undependable<br>Scrupulous / Unscrupulous<br>Persevering / Quitting<br>Fussy / Careless |
| **Emotional Stability** | Calm / Anxious<br>Composed / Excitable<br>Not hypochondriacal / Hypochondriacal<br>Poised / Nervous |
| **Openness** | Intellectual / Unreflective<br>Artistic / Non-artistic<br>Imaginative / Simple<br>Polished / Crude |

**Table 6:** *Personality Traits and Their Positive/Negative Indicators*

Calculating Similarity Using the Embedding Models Using the

**Table 7:** *Personality Position Scores for Each Speaker using Only Positive Indicators*

| Speaker | Extraversion | Agreeableness | Conscientiousness | Emotional Stability | Openness |
|---|---|---|---|---|---|
| Chandler | 0.199479 | 0.208894 | 0.061957 | 0.473858 | 0.101854 |
| Joey | 0.121579 | 0.225151 | 0.083619 | 0.406603 | 0.048025 |
| Rachel | 0.204665 | 0.200387 | 0.037906 | 0.448902 | 0.119116 |
| Monica | 0.207588 | 0.184665 | 0.080646 | 0.432204 | 0.076495 |
| Phoebe | 0.189222 | 0.213814 | 0.084259 | 0.399947 | 0.080105 |
| Ross | 0.196917 | 0.205769 | 0.038643 | 0.496778 | 0.125713 |

**Table 8:** *Personality Position Scores for Each Speaker with Both Positive and Negative Indicators*

| Speaker | Extraversion | Agreeableness | Conscientiousness | Emotional Stability | Openness |
|---|---|---|---|---|---|
| Chandler | 0.129565 | 0.288949 | 0.012685 | 0.482095 | 0.059939 |
| Joey | 0.056017 | 0.265432 | 0.139744 | 0.366637 | 0.053247 |
| Rachel | 0.151934 | 0.276843 | 0.006202 | 0.370955 | 0.134203 |
| Monica | 0.218566 | 0.284546 | 0.099914 | 0.432200 | 0.013158 |
| Phoebe | 0.184464 | 0.286228 | 0.094341 | 0.357724 | 0.057050 |
| Ross | 0.111375 | 0.273657 | 0.034268 | 0.446074 | 0.145537 |

**Table 9:** *Position Scores for Each Personality Trait by Speaker Using extended vocabulary*

| Speaker | Extraversion | Agreeableness | Conscientiousness | Emotional Stability | Openness |
|---|---|---|---|---|---|
| Chandler | 0.081473 | 0.015213 | 0.054863 | 0.056596 | 0.130160 |
| Joey | -0.013197 | 0.030014 | 0.043246 | 0.140287 | 0.050642 |
| Rachel | 0.028611 | 0.052864 | 0.037782 | 0.120380 | 0.132671 |
| Monica | 0.016767 | 0.061303 | 0.027708 | 0.083534 | 0.154021 |
| Phoebe | -0.017234 | 0.046316 | 0.044435 | 0.130028 | 0.117601 |
| Ross | 0.011037 | 0.051259 | -0.010078 | 0.109695 | 0.091352 |

dictionary from the table, we calculated the similarity between the documents (or speaker texts) and each of the personality traits. The calculation is based on the cosine similarity between the document and both the positive and negative indicators for each trait. The following formula was used:

$$similarity(trait) = \sum (\cos(doc, Positive) - \cos(doc, Negative)) \tag{1}$$

This formula calculates the difference in cosine similarity between the document and the positive indicator and the document and the negative indicator for each personality trait. A higher value indicates that the document aligns more closely with the positive side of the trait, while a lower value indicates alignment with the negative side.

### 3.5.1. Running the Analysis

We ran the similarity calculation twice for each document:

1. Using only the Positive Indicators: In this case, the similarity was calculated based solely on the positive indicators for each trait. This approach evaluates how closely the document aligns with the positive aspects of the personality traits.

2. Using Both Positive and Negative Indicators: In this second run, we included both positive and negative indicators to observe how the inclusion of contrasting characteristics affects the similarity scores. By comparing the results from both runs, we were able to assess the impact of the negative indicators on the overall similarity.

The results of our analysis are shown in table 7 and table 8.

## 3.6. Extended Vocabulary for Embeddings

To further experiment with the model, we expanded the vocabulary for each trait to include seven pairs of positive and negative indicators, as shown in table 9 and 10.

## 3.7. Conclusion of the Analysis

Comparing the estimated traits with the dataset labels, we observe that the model performs relatively well. For instance, Joey consistently scores the lowest on Extraversion in both the estimates and labels. Using both positive and negative indicators enhances differentiation, while relying only on positive indicators results in generally higher scores that are close to each other. Expanding the vocabulary further diversifies scores, indicating that carefully selected vocabulary can lead to improved outcomes.

**Table 10:** *Extended Personality Traits and Their Positive/Negative Indicators*

| Trait | Indicator (Positive / Negative) |
|---|---|
| **Extraversion** | Talkative / Quiet |
| | Outgoing / Reserved |
| | Energetic / Lethargic |
| | Expressive / Unexpressive |
| | Assertive / Submissive |
| | Friendly / Aloof |
| | Sociable / Withdrawn |
| **Agreeableness** | Friendly / Hostile |
| | Cooperative / Uncooperative |
| | Compassionate / Apathetic |
| | Forgiving / Vindictive |
| | Trusting / Suspicious |
| | Generous / Selfish |
| | Tolerant / Intolerant |
| **Conscientiousness** | Organized / Disorganized |
| | Dependable / Unreliable |
| | Diligent / Lazy |
| | Thorough / Negligent |
| | Self-disciplined / Impulsive |
| | Detail-oriented / Careless |
| | Goal-focused / Aimless |
| **Emotional Stability** | Calm / Anxious |
| | Stable / Unstable |
| | Confident / Insecure |
| | Relaxed / Stressed |
| | Optimistic / Pessimistic |
| | Resilient / Vulnerable |
| | Content / Discontented |
| **Openness** | Curious / Indifferent |
| | Inventive / Conventional |
| | Insightful / Shallow |
| | Broad-minded / Narrow-minded |
| | Artistic / Unimaginative |
| | Intellectually adventurous / Cautious |
| | Appreciative of diversity / Conforming |

### 3.8. Machine Learning Results

The results of the machine learning on the different variations of the dataset ordered by the mean square error of the testing, goes as follows:

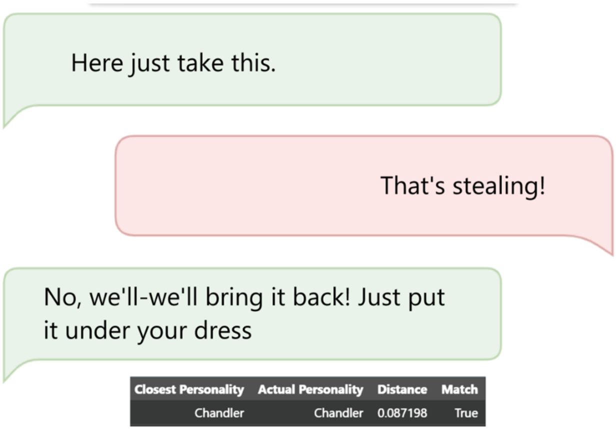
- **trigram**: 0.005757459439337254
- **5000features**: 0.005766735412180424
- **no_emotion**: 0.005812237039208412
- **quadgram**: 0.0058181206695735455
- **no_emotion_or_sentiment**: 0.005820064339786768
- **all_data**: 0.0058225891552865505
- **no_utterance2**: 0.005976484622806311
- **1000features**: 0.006012726575136185
- **bigram**: 0.006078173406422138
- **no_sentiment**: 0.0061368318274617195
- **only_utterance1**: 0.00617994274944067

Where all_data is the baseline dataset where we keep all features of the dataset with a test loss of 0.005823. Removing utterances seemed to degrade the performance the most, as it is removing too much information. However, it is interesting that removing speaker_2 degrades the model, as we are not predicting for speaker_2's personality. Omitting the emotion labels or the sentiment labels either degrades the model or does not make a significant difference, indicating it does not add any significant value for the model.

Limiting the vocabulary to the 1000 most frequent tokens is too limiting as we get a worse performance. However, limiting it to the 5000 most frequent tokens was a better middle ground, where we are not removing too many words, and irrelevant less frequent words are removed, resulting in less noise. Both bigram and quadgram see no improvement, bigram doesn't capture enough of the essence of the sentence, and quadgram creates too much dimensionality, resulting in noise. Trigram seems to strike the middle ground, resulting in an improved model performance.

However, when we try to combine trigram with 5000 token limit, we end up with a worse model again, with a test loss of 0.006021. Indicating that there is a relationship with what information the trigram conveys and the performance. When we remove words, then the trigram becomes noisier.

### 3.9. Abstracting Predicted Personality Scores to Speakers



Here just take this.

That's stealing!

No, we'll-we'll bring it back! Just put it under your dress

| Closest Personality | Actual Personality | Distance | Match |
|---|---|---|---|
| Chandler | Chandler | 0.087198 | True |

**Figure 7:** *Abstraction result*

Abstracting the predicted personality score the closest personality resulted in 310 correct predictions and 992 incorrect predictions. Grouped by the closest personalities the predictions were as follows:

- **Chandler**: 1050
- **Rachel**: 176
- **Ross**: 26
- **Phoebe**: 18
- **Monica**: 16
- **Joey**: 16

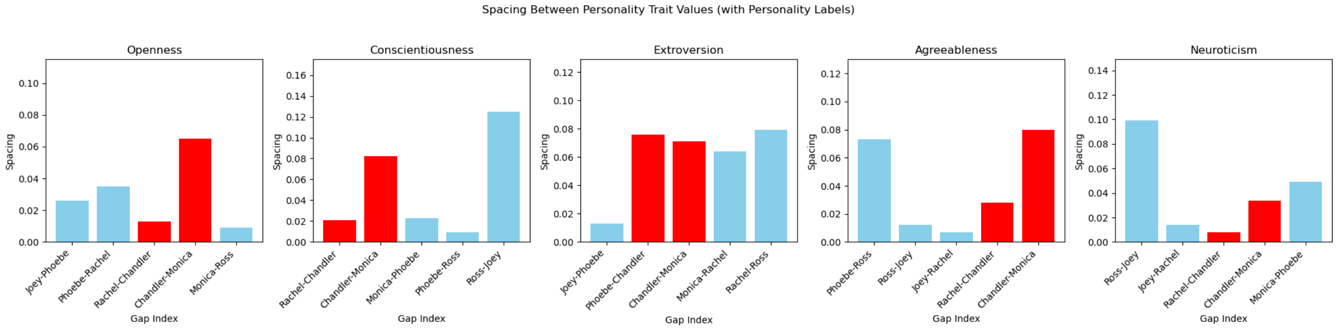


**Figure 8:** *Size of the gap between speakers' personality trait scores, between 0 and 1*

The models abstract to Chandler the majority of the predictions. This does not mean that the model performs poorly, but rather that abstracting five regression scores to a classification is flawed. We'll discuss why in the next chapter.

## 4. Discussion

### 4.1. Abstracting Predicted Personality Scores to Speakers

The results of abstracting the predictions to a speaker were unsatisfactory because this assignment was a regression assignment, and not a classification assignment. We tried to use the five predicted personality scores to determine the speaker, had we instead trained a classification model directly on classifying the speaker, then this would have drastically better results.

The problem with our abstraction arises because when we map 5 scores and calculate the average distance, it will tend to often be closer to a person with more average personality scores. In addition, a trained regression model tends to be on the average-side of wrong. Most predictions predict Chandler, who is more average in all the scores. There is also the talk about how much of the 'score bandwidth' belongs to each person by using proximity. Chandler seems to own a relatively big portion of the average scores.

### 4.2. Further work

This study confirms that we are able to predict and profile personalities based on just conversational data. This would allow for the creation of an artificial intelligence chatbot that can word itself based on personality scores as an input parameter. We could also develop a chatbot that can read the human texter's personality, and in turn respond with the same personality traits, which could increase the user's trust and satisfaction in texting with the chatbot.

## 5. Conclusion

In conclusion, our analysis brings out key insights into personality traits and how well machine learning models can identify them. This analysis revealed that language use, emotional tone, and topic diversity are strongly influenced by personality traits. Key patterns, such as vocabulary size, repetition, and emotional intensity, align with characteristics like extraversion, openness, and agreeableness. While each method offers valuable insights, integrating these approaches provides a more comprehensive understanding of how personality shapes conversational style and emotional expression

Our work with embedding models took this further by comparing how closely speaker texts match both positive and negative trait indicators. Looking at both indicators types compared to only positive one gave us a richer view, though there's room for improvement in how these aspects balance out for better accuracy.

On the machine learning side, the results showed that using trigrams or a 5000-token limit worked best, highlighting the need to find a balance between the amount of data noise. Removing too much information—like certain words or infrequent tokens—reduced performance, showing that a broader set of data helps accuracy.

Overall, these findings reveal some of the strengths and challenges in personality prediction. While we made good progress in identifying patterns and optimizing model performance, there's still room to refine the way we preprocess the data to best match the personality traits.

## References

[Ble03] Blei D. M. N. A. Y. . J. M. I.: Latent dirichlet allocation. *Journal of Machine Learning Research 3*, 1 (Jan. 2003), 993–1022. 2

[JYS*22] Jang J., Yoon S., Son G., Kang M., Choeh J. Y., Choi K.-H.: Predicting personality and psychological distress using natural language processing: A study protocol. *Frontiers in Psychology 13* (2022). URL: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2022.865541, doi:10.3389/fpsyg.2022.865541. 12

[McC99] McCrae R. R. . C. P. T.: A five-factor theory of personality. *Handbook of Personality: Theory and Research (2nd ed., pp. 139-153). Guilford Press. 2*, 1 (Jan. 1999), 1–34. 3

[MPGC17] Majumder N., Poria S., Gelbukh A., Cambria E.: Deep learning-based document modeling for personality detection from text. *IEEE Intelligent Systems 32* (03 2017), 74–79. doi:10.1109/MIS.2017.23. 4

[PBCP20] Pradhan T., Bhansali R., Chandnani D., Pangaonkar A.: Analysis of personality traits using natural language processing and deep learning. In *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)* (2020), pp. 457–461. doi:10.1109/ICIRCA48905.2020.9183090. 4