International Journal of Applied Science; Vol. 8, No. 4; 2025
ISSN 2576-7240 E-ISSN 2576-7259
https://doi.org/10.30560/ijas.v8n4p54

Sentiment and Personality Shifts in Code-Switching: A Transformer-Based NLP Analysis

Haixiang Gong[1]

[1] Phillips Academy Andover, USA

Correspondence: Haixiang Gong, Phillips Academy Andover, Andover, MA, USA.

Received: November 12, 2025; Accepted: November 23, 2025; Published: November 24, 2025

Abstract

Cultural frame switching posits that bilinguals adjust their personalities according to the cultural context of the language they speak. Prior research explored this through psycholinguistic surveys and self reported data. With advances in Natural Language Processing and transformer architectures, this study reexamines cultural frame switching using machine learning methods. We computationally test whether bilingual personality expression varies across code-switched and monolingual speech, using sentiment deltas as proxies for personality shifts. Leveraging the SEAME corpus (110,145 utterances:57,052 code-switched,28,655 English monolingual, 24,438 Mandarin monolingual, 156 speakers), five transformer-based sentiment classifiers were applied to each entry. All models detected significant sentiment differences across language bins, with three of four sentiment-tuned models showing consistent relative orderings between code-switched, English, and Mandarin speech, while one outlier model produced inconsistent results. Statistical analysis supports the psychological theory of cultural frame switching.

Keywords: cultural frame switching, bilingual personality, SEAME corpus, code-switching, trans-former models

1. Introduction

Code-switching, or alternating between two or more languages in a single utterance, is an important aspect of bilingual life and is widespread in bilingual communities. This phenomenon, however, has only been examined primarily through sociolinguistic and pragmatic lenses (Liu, 2019; Zhang, 2019). Psycholinguistic studies have examined the fundamental cognitive and emotional frameworks that explain these behaviors. Some pro-pose that language choice may affect the expression of certain personality traits, formalized as *cultural frame switching* (Ramírez-Esparza et al., 2006). This study builds upon this framework by computationally analyzing this theory through analyzing personality shift through the proxy of sentiment in transcribed natural speech.

This study asks whether these personality shifts are reflected in sentiment polarity between code-switched (CS), English monolingual (EM), and Mandarin monolingual (MM) utterances. We analyze whether multilingual transformer-based NLP models can detect such shifts in sentiment when applied to real-world, code-switched speech data. Through this approach, we aim to answer the following questions: Does sentiment polarity or intensity consistently differ between code-switched sentences compared to non-code-switched sentences from the same individual? Do multilingual sentiment analysis models consistently capture differences in sentiment polarity or intensity between English-only and Mandarin-only utterances? Do current multilingual sentiment analysis models detect the nuances present in code-switched languages?

2. Literature

Bilingual research, specifically in personality expression, indicates consistently that a speaker's language choice elicits language-distinct emotional and thought patterns, manifesting in different personality expressions. This phenomenon was first explored by Susan Ervin in French-English bilinguals through thematic apperception stories, where study participants produced narrative in response to ambiguous images (Ervin, 1964). The study found that participant's narratives in French revolved around themes of guilt and authority, while their English ones emphasized aggression and achievement. Ervin posited that bilinguals may experience different "modes of perception" under different lingual contexts. This early work established the idea that bilingual personality expression is fluid, with language acting as a situational angle for one's expression. Michèle Koven's interviews corroborated this observation: her study found that speakers ex-pressed different personal qualities by language when telling stories (Koven, 1998). Individuals appeared more assertive and "foul mouthed" in French while

seeming more polite and modest in Portuguese when recounting identical anecdotes. Koven interpreted this as bilinguals aligning them-selves with the cultural context of each language.

Ramírez-Esparza et al. formalized this idea when exploring personality traits in Spanish-English bilinguals through the Big Five frame-work (Ramírez-Esparza et al., 2006). Participants took personality assessments in each language and showed significant differences between them. In English, participants rated themselves higher in Extraversion, Agreeableness, and Conscientiousness compared to Spanish, agreeing with established cultural norms. The authors described this as an instance of cultural frame switching, where bicultural-bilinguals alter their personality expression to accommodate the cultural context of a certain language. Cultural frame switching theory posits participants' underlying personalities do not change; instead, each language only draws out specific aspects of their personality, fitting the cultural context of each language. These findings have also been corroborated by other researchers (Chen and Bond, 2010) and even replicated with other bilingual populations (Dewaele, 2013). The results observed in all these studies indicate that bilinguals adjust to the distinct cultural norms associated with each language environment.

Sentiment analysis offers a robust computational proxy for personality when analyzing shifts. Mairesse et al. (2007) demonstrated the inference of the Big Five traits from speech transcripts, while Li et al. (2022) found robust correlations between personality dimensions and affective language markers, notably neuroticism with negative affect (on sentence sentiment) and extraversion/agreeableness with positive affect (Mairesse et al., 2007; Li et al., 2022). These studies establish ground truth for utilizing sentiment as a way to detect shifts in expressed personality.

Transformer architectures, utilizing self-attention mechanisms for long range dependency handling, allow for more nuanced sentiment analysis in text (Vaswani et al., 2017).Con-neau et al. (2020) demonstrated transformers' effectiveness in multilingual text, providing the technical foundation for transformer models to consistently detect sentiment variations in multilingual utterances (Conneau et al., 2020).

Analysis of the "mixed" variable may also relate to postcolonial theorist Homi K. Bhabha's concept of a third space, where cultural hybridity is not a fusion of two distinct identities, but a product of them. In the context of code-switching, bilinguals may use codeswitching not simply as middle ground between their two lingual poles, but instead construct a distinct "third" identity situated perhaps outside the two ends of the lingual poles (Bhabha, 1994).

3. Method and Experiments

3.1 Corpus

The Southeast Asia Mandarin–English (SEAME) corpus, an annotated spontaneous speech database of 156 speakers, was the source of utterances for sentiment analysis. The corpus comprises roughly 120 hours of recordings or 110,145 utterances, with a distribution of 28,655 EM, 24,438 MM, and 57,052 CS utterances. This dataset provides a solid foundation for our analysis (Lyu et al., 2015).

The SEAME corpus was recorded in Singapore and Malaysia in 2015 in two speaking styles: con-versational dialogues and interviews. The conversational sessions were recorded at Nanyang Tech-nological University (in Singapore), and interview sessions were recorded at both NTU and Universiti- Sains Malaysia. Speakers in the SEAME corpus were between the age of 19-33 (~50% female) and were one of two nationalities (~63% Singaporean,~37% Malaysian). English is the first language for Chinese Singaporeans (English L1, Mandarin L2), whereas Mandarin is the primary home language for most Chinese Malaysian speakers (English L2, Mandarin L1). These demographic differences provide an interesting comparison, although all participants are bilingual in English and Mandarin.

3.2 Method

We segmented the corpus into three subsets: utterances spoken entirely in English (EM), entirely in Mandarin (MM), and mixed utterances (CS) containing both English and Mandarin or, in other words, code-switched instances. Sentiment models were used to assign sentiment scores to each utterance in these bins. By comparing sentiment scores between these subsets, we aim to capture and quantify differences in emotional expression during code-switching.

We employed several Python scripts in our pipeline. Each file accepts a data file and outputs a processed version. First, parse.py converted raw transcripts into structured utterances (excluding tags like <v-noise> and <unk>). Then, process.py ran each NLP model on each utterance and recorded the sentiment predictions. We then applied normalize.py to flatten each model's output to a common polarity scale (using the formulas below). Finally, by-person.py and other scripts plotted histograms, or fitted linear models. (Details are available in our repository) (Gong, 2025)

3.3 Models

Using five models, our methodology triangulates each utterance's sentiment via classifiers with diverse backgrounds. Each model has distinct training backgrounds, but all can process CS inputs due to multilingual pretraining. Thusly, code-switched utterances can be fed directly without additional processing. We recorded each model's sentiment score per utterance. General agreement between models on a sentiment trend would indicate a likely pattern, while differences could suggest insensitivity to bilingual nuances for certain models. This strategy allows us to confidently examine differences in model behavior through shared consensus. Four of the five classification models assigned labels with corresponding confidence scores (which sum to 1.00 across all labels). Compression into a continuous sentiment score was performed with the following formula:

$$S = \sum_{i=1}^{k} W_i S_i$$

where:

- S = total sentiment score,
- S_i = score for category i,
- W_i = weight assigned to category i ,
- k = number of sentiment categories.

The two specific formulas used to compress label probabilities into a single score across models are as follows:

$$\text{Sentiment Score} = (\text{Negative}) \times (-2) + (\text{Positive}) \times 2 ,$$

$$\text{Sentiment Score} = (\text{Very Negative}) \times (-2) + (\text{Negative}) \times (-1) + (\text{Positive}) \times 2 + (\text{Very Positive}) \times 2 .$$

Model A (Multilingual DeBERTa Sentiment)

The first model (agentlans/ mdeberta-v3-base-sentiment) is a fine-tuned Multilingual DeBERTa (DeBERTa V3 base). Model A outputs a continuous sentiment score between −1 and 1. It was trained on ~48,000 English sentences from a Tatoeba corpus and their translations with sentiment labels. This training generates a model which outputs a single sentiment value (positive = positive sentiment, negative = negative sentiment, zero = neutral) when given text. Model A's direct regression output meant that no score flattening was required.

Model B (XLM-RoBERTa Large XNLI)

The second model (joeddav/ xlm-roberta-large-xnli) uses Facebook's XLM-RoBERTa Large architecture and was fine-tuned on the Cross-lingual NLI (XNLI) dataset. Model B is a robust multilingual model trained on 2.5 TB of text across 100 languages. The XNLI fine-tuning process involves determining whether a given "hypothesis" is entailed by, neutral to, or contradicted by a given "premise" across languages. Although not designed for sentiment, we use Model B in a zero-shot manner: the utterance is given as the premise, and a templated sentiment statement (e.g., "The sentiment of this phrase is {positive/negative/neutral}.") as the hypothesis. The model then infers the most likely sentiment label. Including this NLI-based model provides a baseline to compare zero-shot performance against dedicated sentiment models.

Model C (DistilBERT Multilingual Sen-timent)

The third model (tabularisai/ multilingual-sentiment-analysis) is built on a DistilBERT-base-multilingual-cased architecture. DistilBERT uses model distillation to maintain ~97% of BERT's performance with ~40% fewer parameters (Sanh et al., 2019). The Tabularis model was trained exclusively on synthetic sentiment data generated by large language models, predicting sentiment in five classes (Very Negative, Negative, Neutral, Positive, Very Positive). This allows more nuanced scores when flattened. Despite the unconventional training data, the model achieved high validation accuracy (~93% within one class), suggesting it learned multilingual sentiment effectively. We include Model C for its efficiency and to observe how an LLM-trained sentiment classifier performs on real code-switched speech.

Model D (ModernBERT-large Multilin-gual Sentiment)

The fourth model (clapAI/modernBERT-large-multilingual-sentiment) represents in this study the ModernBERT archi-tecture, a 2024 reimplementation of the original BERT with modern enhancements (Warner et al., 2024). Model D was fine-tuned on ClapAI's Multi-lingualSentiment dataset (~3.9M examples from 17 languages, including Mandarin and English) with three sentiment labels (positive, neutral, negative). Model D is used as a state-of-the-art baseline to see how a cutting-edge multilingual sentiment model handles code-switched input.

Model E (CardiffNLP XLM-RoBERTa Base)

The fifth model (cardiffnlp/ xlm-roberta-base-sentiment-multilingual) uses the XLM-RoBERTa architecture (same as Model B) but was fine-tuned specifically for sentiment on the Tweet Sentiment Multilingual dataset (Barbieri et al., 2022). Although the fine-tuning data did not include Mandarin, the multilingual pretraining of the XLM-RoBERTa architecture compensates for this. We included Model E to directly compare with Model B: both are XLM-RoBERTa models, but Model E has sentiment-specific fine-tuning. Comparison between this model and Model B highlights the importance of task-specific training for capturing code-switched sentiment.

This study did not involve human subject participation. All data were drawn from the publicly available SEAME corpus (Lyu et al., 2015). No personally identifiable information was accessed or analyzed; therefore, IRB approval was not required.

Table 1. Summary of models used in sentiment comparison experiments

Model	Architecture	Reasoning
Model A	Multilingual DeBERTa – sentiment re-gression model (no flattening needed)	Serves as a comparison between regression and classification models.
Model B	XLM-RoBERTa Large XNLI – zero-shot NLI-based model	Not trained for sentiment; provides a baseline vs. task-specific models.
Model C	DistilBERT Multilingual Sentiment – distilled BERT model	Tests an LLM-trained sentiment classifier on real code-switched data.
Model D	ModernBERT-large Multilingual Senti-ment – fine-tuned ModernBERT	Represents a cutting-edge model; serves as a strong baseline for comparison.
Model E	XLM-RoBERTa Base – same architec-ture as B, fine-tuned for sentiment	Enables direct comparison to the XNLI-based Model B (effect of sentiment fine-tuning).

4. Results

Our results show clear differences in sentiment across the three language categories (EM, MM, CS). Overall, EM utterances tend to receive more positive sentiment scores, MM utterances more negative scores, and CS utterances fall in between.

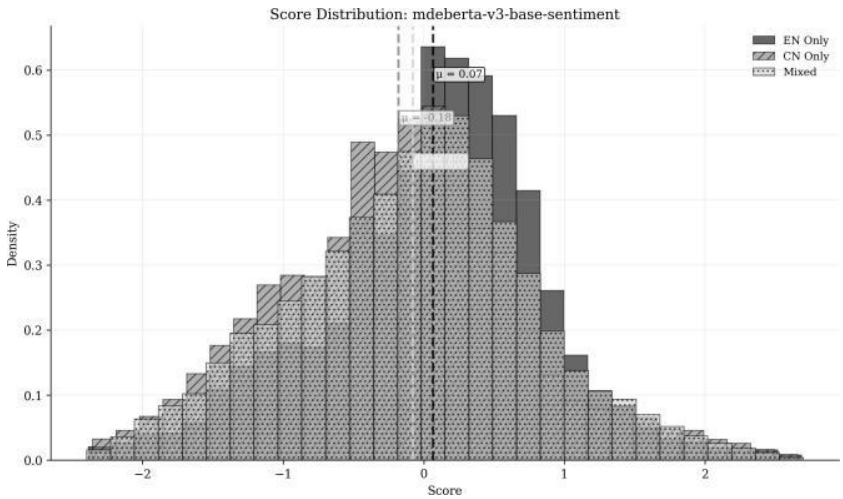

Figure 1. Histogram of sentiment scores for Model A

This pattern was consistent in four of the five models (with the exception of one outlier, likely due to its training objective), though the exact CS placement varied. Below are sentiment histograms for each model, highlighting mean polarity by language and notable distribution features, followed by significance testing.

4.1 Means

Table 2. Mean sentiment scores (by utterance) by model and language status

Model	EN Only (n=23,775)	CN Only (n=21,162)	Mixed (n=54,108)
Model A	0.0664	-0.1804	-0.0788
Model B	0.3128	0.4693	0.3869
Model C	0.0398	-0.1476	-0.0843
Model D	0.0850	-0.2148	-0.3198
Model E	-0.0815	-0.1909	-0.1459

4.2 Histogram

Model A (Multilingual DeBERTa)

Figure 1 illustrates Model A's sentiment distributions. EM (blue) utterances skew slightly positive (mean ∼ 0.11); MM (green) utterances center lower (mean∼ −0.19); CS (red) utterances are intermediate (∼ −0.09). The EM distribution is approximately normal, whereas MM and CS distributions are right-skewed.

Model A shows a standard ordering of sentiment means of Mandarin < CS < English. The agreement of Model A with others verifies the method of sentiment label compression used.

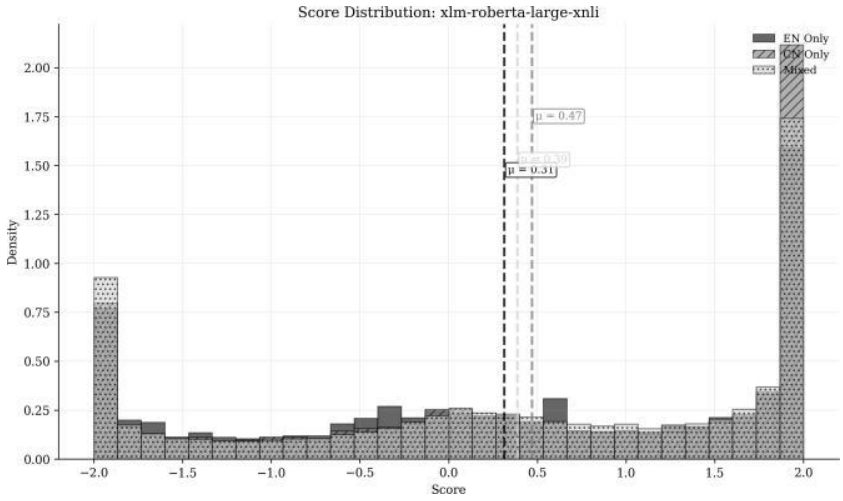

Figure 2. Histogram of sentiment scores for Model B

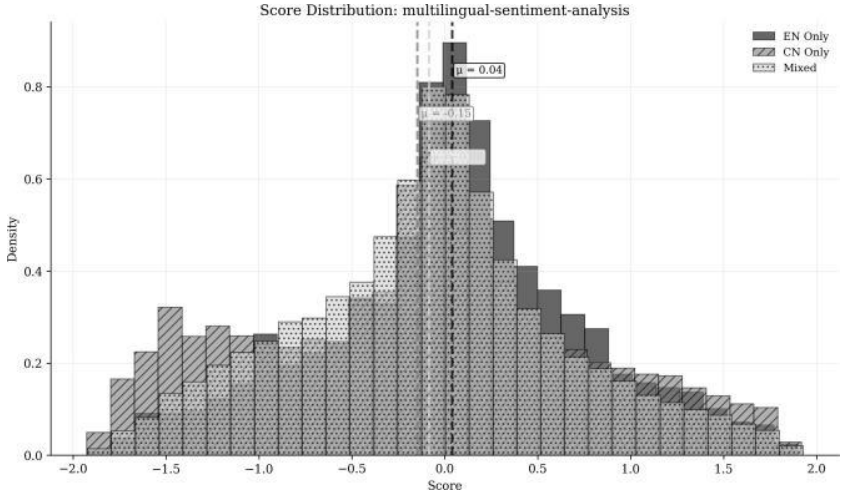

Figure 3. Histogram of sentiment scores for Model C

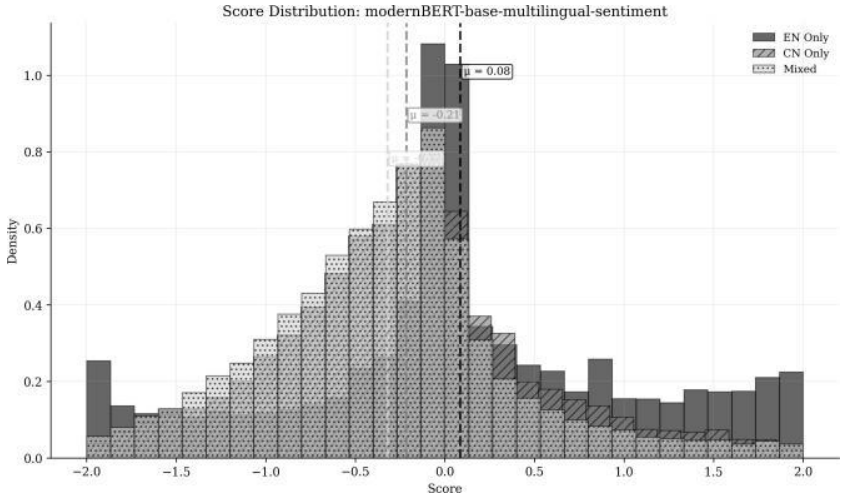

Figure 4. Histogram of sentiment scores for Model D

Model B (XLM-RoBERTa XNLI)

Figure 2 shows Model B's sentiment distributions. EM (blue), MM (green), and CS (red) curves nearly overlap, with means ~ 0.54 (EM), ~ 0.56 (MM),~ 0.47 (CS).

Model B shows no clear sentiment polarity differences between the language categories. As illustrated in Figure 2, EM, MM, and CS utterances have almost identical sentiment distributions (means ~ 0.51–0.56). The zero-shot NLI classifier essentially assigns similar sentiment scores regard-less of language. In other words, Model B treats utterances in English and Mandarin almost equally in sentiment. Likely explanations include Model B's training for zero-shot usage; its training objective was not tuned to detect affective cues. Thus, the absence of a detected sentiment shift may reflect Model B's limitations rather than evidence against a real bilingual sentiment difference.

Model C (DistilBERT Multilingual)

Figure 3 shows Model C's distributions. EM (blue) utterances have a near-neutral mean (~ 0.01) and a roughly symmetric shape. MM (green) utterances are right-skewed (mean ~ −0.10). CS (red) utterances show an intermediate mean (~ −0.08) with a slight right skew.

Model C exhibits clear polarity differences across English, Mandarin, and mixed inputs, and follows the consistent pattern in other models. This model's Mandarin < CS < English ordering of mean sentiment is well-separated for Model C. Model C suggests each language context produces a distinct sentiment distribution. Its nuanced five-level sentiment scale likely contributes to capturing these differences.

Model D (ModernBERT-large)

Figure 4 shows Model D's distributions. EM (blue) and MM (green) utterances center near neutral (means ≈ 0.17 and −0.17). CS (red) utterances skew slightly negative (mean ≈ −0.30), sitting below both mono-lingual groups.

Model D detects sentiment differences across all lingual bins, but with a divergent pattern. English and Mandarin utterances receive similar near-neutral scores (English ≈ 0.17, Mandarin ≈ −0.17). Uniquely, the CS mean is more negative than even the Mandarin mean (≈ −0.30). For Model D, then, the ranking is English > Mandarin > CS in positivity, rather than CS falling between English and Mandarin. The distribution shapes differ as well: the MM and CS curves have thin tails, whereas the EM curve has thicker tails (more frequent extremes), which may partly explain Model D's departure from the others. This departure may be a manifestation of more confidence in English from a more robust English-handling architecture, resulting in a higher rate of extreme predictions. While the rank order (EM highest, CS lowest) is notable, the consistent EM vs. MM difference aligns with other models, indicating Model D's mixed-input behavior is idiosyncratic. The divergent CS position could be a manifestation of a general gravitation, albeit to a drastic degree, of the CS mean towards the MM mean, which is evident in the three other sentiment-tuned models. Additionally, the drastic difference between MM and EM averages further supports the theory of *cultural frame switching*: despite the unusual CS mean placement, the EM and MM means remain in similar directions.

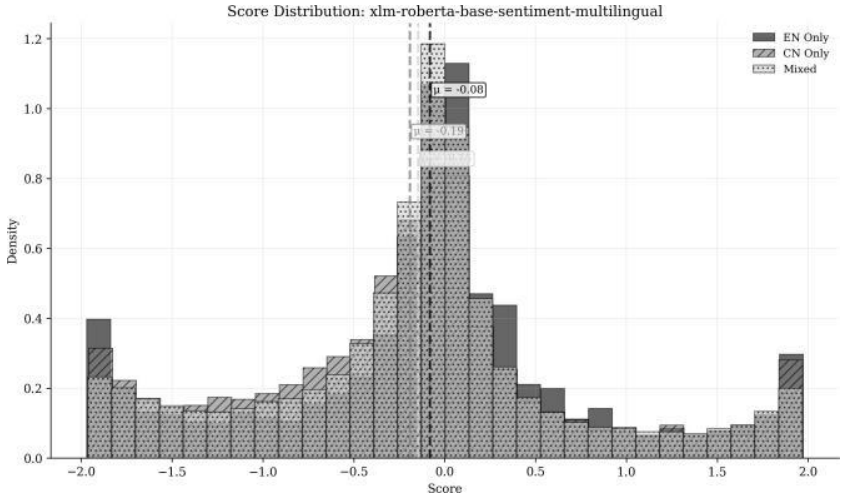

Figure 5. Histogram of sentiment scores for Model E

Model E (XLM-RoBERTa Base)

Figure 5 shows Model E's distributions. MM (green) utterances have a slight right skew (mean ∼ −0.21). CS (orange) utterances skew somewhat rightward (mean ∼ −0.15). EM (blue) utterances center near neutral (mean ∼ −0.02).

Model E shows sentiment patterns similar to the other tuned models. EM utterances have the highest (least negative) average sentiment (∼ −0.02), MM utterances the lowest (∼ −0.21), and CS utterances fall in between (∼ −0.15). These results follow the general trend of EM > CS > MM for sentiment means. This aligns with cultural frame switching, suggesting that this model too detects shifts in sentiment when a speaker switches languages. Notably, because Model E was fine tuned for multilingual sentiment, its ability to assign consistent directionality between EM and MM is not an artifact of one architecture, instead it appears across diverse models with the training task of sentiment detection (contrasting with the XNLI-based Model B, which lacked sentiment fine-tuning).

4.3 By-Person Analysis

We confirmed the observed trends at the individual speaker level. Figure 6 shows histograms of the mean sentiment difference for each speaker between language conditions, by model. These speaker-level patterns mirror the model-level results.

4.4 Linear Models

Table 3. Regression coefficients across models. Significance codes: *** p < 0.001, ** p < 0.01, * p < 0.05, ·p < 0.1. Dummy coding: Mixed 0 = EM, Mixed 1 = MM, Mixed 2 = CS

Variable	mDeBERTa-v3-base	XLM-R large-XNLI	ModernBERT-base-multi	Multilingual-SA	XLM-R base-multi
Intercept	0.002	0.164***	0.093***	-0.020	-0.068**
C(location)[T.U]	0.101**	0.205***	-0.070*	0.019	0.033
C(gender_clean)[T.M]	0.037**	0.060**	0.004	-0.002	0.020
C(nationality)[T.B]	0.071***	0.145***	0.004	0.069***	-0.001
C(nationality)[T.F]	-0.025	-0.243*	-0.210***	0.025	-0.223***
mixed_1	-0.272***	0.144***	-0.277***	-0.173***	-0.141***
mixed_2	-0.172***	0.046***	-0.393***	-0.112***	-0.095***

In the four sentiment-tuned models, most individual's EM utterances were rated more positively than their MM utterances. The histogram of each model's English–Mandarin differences is strongly skewed, with the vast majority of speakers showing English > Mandarin sentiment (i.e., a negative value for Mandarin minus English). Only a few speakers cluster near zero or show a reverse trend, indicating the English-positive/Mandarin-negative pattern holds for most speakers. The sole clear exception is Model B, whose speaker-level difference distribution is centered near zero (as expected from its overall null result). Similar trends appear for differences involving CS

utterances: in Models A, C, and E, each speaker's CS sentiment lies between their English and Mandarin means (often closer to their Mandarin sentiment), whereas Model D some-times shows a speaker's CS sentiment even lower than their Mandarin sentiment (reflecting Model D's overall pattern). In sum, the speaker-level analysis confirms that the sentiment gaps observed in aggregate are not driven by outliers but are pervasive across participants.

We also ran linear mixed-effects regression models for each sentiment model's scores. In each model, language context (English vs. Mandarin vs. Mixed) was a fixed effect (with dummy coding), with speaker-level controls (location, gender, nationality) and a random intercept for speaker to ac-count for individual baselines. These mixed-effects results add nuance to the histogram findings. Table 3 displays coefficients and significance for each factor using the following formula for linear regression prediction:

$$\text{Sentiment}_{ij} \sim \text{C(location)} + \text{C(gender\_clean)} + \text{C(nationality)} + \text{mixed\_1} + \text{mixed\_2}$$

For Models A, C, D, and E, language context remains a significant predictor of sentiment even after controlling for speaker and demographics. For example, in Model A, the difference in mean sentiment from EM to MM is roughly -0.27 (normalized units), and CS differs by -0.17 from EM. Both effects are significant ($|p| > 24$, $p \ll 0.001$). Similarly large negative coefficients for Mandarin and mixed (relative to English baseline) appear in the Tabularis, ModernBERT, and XLM-R Base models, with extremely small p-values, confirming that those models assign lower sentiment to Mandarin and code-switched speech than to English, all else equal. Notably, the Mandarin effect is generally larger than the mixed effect, reflecting that while mixed utterances are negatively biased, they usu-ally aren't as low as purely Mandarin ones (again echoing the histogram means). These regressions indicate the language-based sentiment shift is robust across speakers.

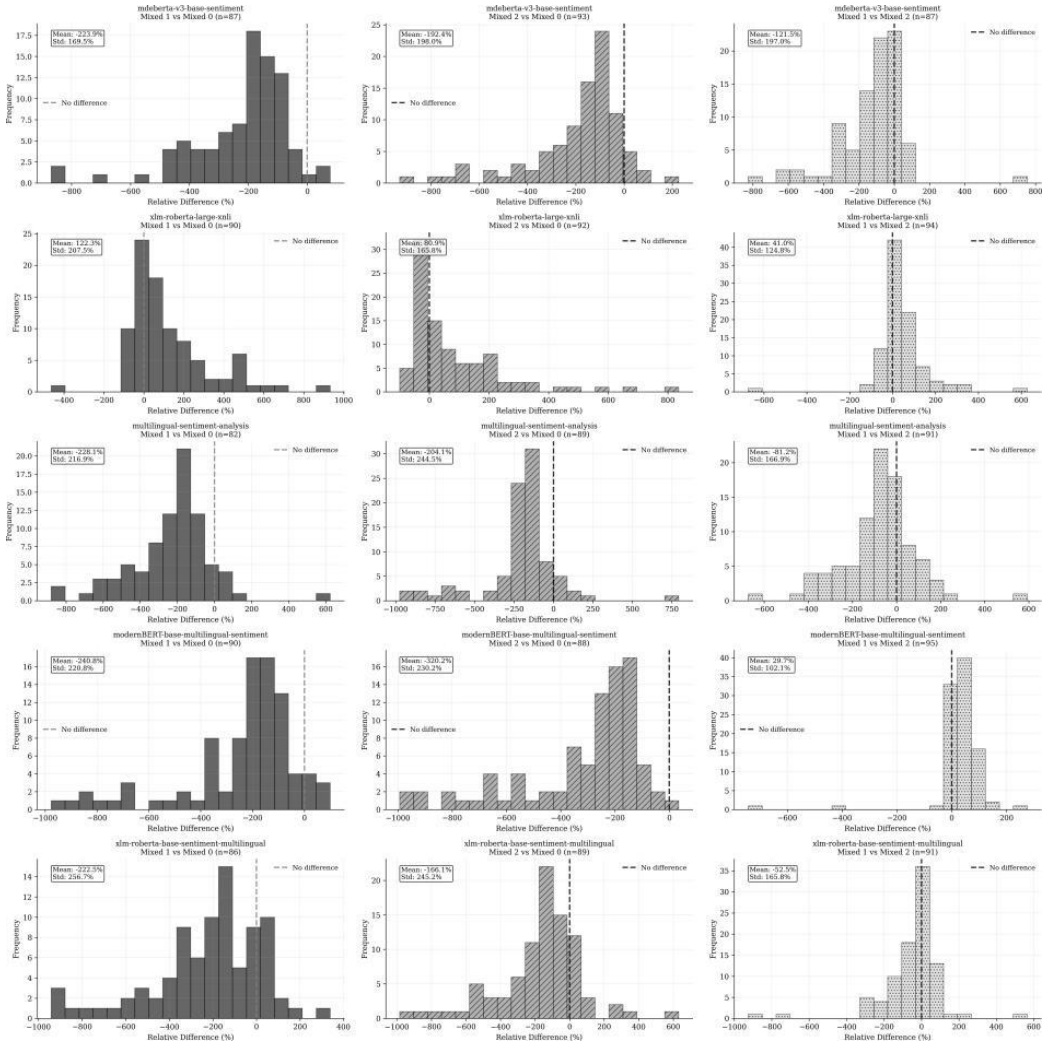

Figure 6. Histogram of by-person sentiment averages for EM, MM, and CS bins.

5. Discussion

5.1 Research Questions

Our results indicate bilingual speakers' sentiment polarity shifts with language choice. Across four of the five models, EM utterances were rated more positively, MM utterances more negatively, and CS speech fell in between while skewing towards MM utterances. For example, Model C found English sentences near neutral on average (~ 0.01) versus negative for Mandarin (~ -0.10), with mixed utterances intermediate (~ -0.08). These differences were often statistically significant, shown in the GLM regression done (all $p < 0.001$). However, aside from p-values, Model B was an outlier in most other aspects such as sentiment means by per-son/phrase and directionality of EM, MM, and CS sentiment affect in the GLM regression. This zero-shot model failed to detect the nuanced sentiment shifts, likely due to its NLI training objective, rather than the sentiment objective of all other models. In contrast, the sentiment-tuned models consistently detected the pattern. Notably, one model (Model D) produced an outlier ordering where code-switched inputs were scored slightly more negative than even MM ones (means ~ -0.30 vs. -0.17), yet it still distinguished CS sentiment as significantly different from both monolingual modes.

Interpreting the intermediate sentiment of mixed utterances through a postcolonial lens of hybridity, our results indicate that the statistical significance of differences between these can support the theory to a degree. However, the directionality of the sentiment mean of the CS bin, positioned between the poles of EM and MM, consistent in three of the five models (in more accurate words, since Model B was an outlier in most aspects, three of four models with the objective of sentiment analysis agree).

While an intermediate "mixed" sentiment average of CS does not necessarily discount cultural hybridity, due to the singular dimensionality of sentiment, the ordering of sentiment means of our results does not support the concept of cultural hybridity. Of course, since the models utilized in this study are black-box models, we cannot ascertain whether the low-level transformer architecture, consistent across models, contribute to code-switched phrases being treated as blended.

5.2 General Trends and Proximity of CS and MM Means

Across models, we observe consistent effects of code-switching on sentiment alongside some model-specific differences. The CS results indicate that mixed utterances have a distinct sentiment profile. While not all models showed the same ordering, a few generalizations hold. First, mixed utterances were always significantly different in sentiment from both EM and MM utterances. Second, in most sentiment-tuned models (e.g., Model A, Model C), mixed speech sentiment skewed closer to the Mandarin baseline, suggesting that the presence of Mandarin in an utterance tends to pull sentiment in a negative direction. Notably, Model E, a robust multilingual sentiment model, followed this pattern as well: its average CS sentiment (-0.15) was closer to the MM mean (-0.21) than the EM mean (-0.02). The consistency of this result across four of five models suggests the negative bias of the CS bin is not an artifact of any single model. In-stead, it likely reflects an underlying reality: code-switched expressions are interpreted (by these models, at least) in line with the culturally influenced sentiment of Mandarin speech.

Despite general consistency, certain models deviate. Model D exaggerated the trend by scoring mixed utterances even lower than Mandarin utterances. This inversion (English > Mandarin> CS) is an outlier, likely due to Model D's training or architecture. For instance, its English predictions showed unusually broad tails (high-confidence extremes absent for Mandarin or mixed inputs). Despite its unusual ordering, Model D still distinguished CS sentiment from both monolingual conditions (with significant pairwise differences). On the other hand, the zero-shot XLM-RoBERTa model (Model B) struggled to differentiate sentiment by language: its English and Chinese means were virtually identical, and only the mixed vs. monolingual comparisons were significant. This implies that Model B, which is trained for NLI and not fine-tuned for sentiment, lacked the sensitivity to detect cultural frame switching in sentiment. Moreover, Model B's average scores were unusually high (around 0.4–0.5, whereas others were near 0), indicating this model's sentiment outputs are unreliable for our task. In contrast, Model E's strong adherence to the expected pattern reinforces that the mixed-language sentiment gap is likely genuine. Because Model E was fine-tuned for sentiment, its ability to capture the mixed-language shift suggests that code-switched utterances truly carry a sentiment "accent" of the Mandarin context.

The tail behaviors of the two XLM-RoBERTa models (B and E) suggest architectural quirks. Both show an unusual trimodal distribution (spikes at each end and near neutral), perhaps an indication that XLM-RoBERTa tends be confidence with extreme predictions, or perhaps an indicator of training data including the SEAME.

In summary, a consistent pattern emerges within a majority of models: EM means skew more positive, MM means skew more negative, and CS means occupy an intermediate position (often closer to Mandarin). The CS bin's

distinct position, isolated from both English and Chinese, reflects the features of code-switching. It appears that when speakers blend English and Mandarin, sentiment tends to gravitate (while still remaining distinct from the MM mean) toward the Chinese/Mandarin side, whether due to genuine linguistic/cultural framing or model bias. The overall convergence of results from a majority of models (A, C, D, E) supports the theory of cultural frame switching.

At the same time, our findings show that model architecture and training data matter greatly, best exemplified by Model B, which was not fine-tuned for sentiment, deviating significantly from other sentiment-tuned models. Nonetheless, future re-search, perhaps comparing model outputs to human judgments of code-switched sentiment, would help determine whether mixed utterances are truly more negative or if our models exhibit a bias due to Mandarin content.

5.3 Limitations and Future Work

A notable limitation of this study is our reliance on the single-dimensional metric of sentiment. This approach cannot identify which specific personality traits are changing. Our approach also relies on the assumption that all personality traits affect sentiment to the same degree, which may not be the case. For example, collapsing multiple personality dimensions into a single sentiment axis risks traits unrelated to sentiment (geometrically orthogonal) giving minimal sentiment affect when projected.

In response to this, we propose "personality tensors." By applying psycholinguistic frameworks like the Big Five, one could represent personality as a vector, or even tensor, of distinct traits. This approach allows for quantitative comparisons in an n-dimensional trait space. Instead of relying on sentiment deltas and its innate information loss, one could measure personality gradients, bypassing the need for a proxy. Our results demonstrated the importance of appropriate model tasks through the poor performance of models not fine-tuned for sentiment (Model B). Therefore, future work on NLP models fine-tuned to output personality vectors from code-switched inputs could yield more distinct and targeted insights.

5.4 Conclusion

Our study shows that sentiment polarity in bilingual speech varies systematically with language con-text. Using five multilingual transformer models on the SEAME code-switching corpus, we found that English-only utterances are rated more positively, Mandarin-only utterances more negatively, and mixed-language utterances fall in between, though consistently gravitating toward the Mandarin-only utterance means. This pattern was observed by most sentiment fine-tuned models, but a general zero-shot model (Model B) was less effective at detecting these nuances, emphasizing the importance of fine tuning for specific tasks. The results from the sentiment-tuned models align with the concept of *cultural frame switching*: bilingual speakers may adopt different emotional or personality profiles depending on the language they're speaking. Code-switched speech appears to represent a blended emotional register, combining aspects of each language's affective profile, though they tend to skew towards the Mandarin-only mean in the case of English-Mandarin bilinguals. However, when analyzing the concept of the "third space," our results found that three of the four sentiment-tuned models assigned an intermediate value for the CS mean, between the EM and MM poles. These results highlight the study's major limitation of a single dimensional variable (sentiment) being analyzed, meaning no concrete support or refutation could be extrapolated from these results. Our findings demonstrate the potential for transformer-based NLP to uncover subtle psycholinguistic patterns in transcribed speech previously difficult to analyze consistently. These insights have broader implications. With the importance of task-specific fine tuning established clearly by the non sentiment-tuned model's (Model B) poor performance, our findings call for opportunities to refine sentiment models in code-switching specific contexts and perhaps even for NLP models with high dimensional personality tensor outputs.

Acknowledgments

We thank the creators of the SEAME corpus for creating the main dataset for this code-switching research, as well as the developers of the multi-lingual transformer models used in this study. We also thank the open-source NLP community, specifically the Hugging Face community and its contributors to multilingual sentiment analysis datasets and models, for enabling reproducible computational experiments.

Huge thanks to Dr. Andrew Nevins for mentoring me through this process. He walked me through all the parts of conducting and writing research, and provided insightful feedback for my work.

References

[1] Barbieri, F., Espinosa-Anke, L., & Camacho-Collados, J. (2022). XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth International Conference on Language Resources and Evaluation (LREC 2022)* (pp. 6750–6759). European Language Resources

Association.

[2] Bhabha, H. K. (1994). *The location of culture*. Routledge.

[3] Chen, S. X., & Bond, M. H. (2010). Two languages, two personalities? Examining language effects on the expression of personality in a bilingual context. *Personality and Social Psychology Bulletin*, *36*(11), 1514–1528. https://doi.org/10.1177/0146167210385360

[4] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Joulin, A., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)* (pp. 8440–8451). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.747

[5] Dewaele, S. (2013). Multilinguals' perceptions of feeling different when switching languages. *Journal of Multilingual and Multicultural Development*, *34*(2), 107–120. https://doi.org/10.1080/01434632.2012.712133

[6] Ervin, S. M. (1964). Language and TAT content in bilinguals. *Journal of Abnormal and Social Psychology*, *68*(5), 500–507. https://doi.org/10.1037/h0044803

[7] Gong, H. (2025). *NLP for multilingual personality research*.

[8] Koven, M. (1998). Two languages in the self/the self in two languages: French-Portuguese bilinguals' verbal enactments and experiences of self in narrative discourse. *Ethos*, *26*(4), 410–455. https://doi.org/10.1525/eth.1998.26.4.410

[9] Li, Y., Kazameini, A., Mehta, Y., & Cambria, E. (2022). Multitask learning for emotion and personality traits detection. *Neurocomputing*, *493*, 340–350. https://doi.org/10.1016/j.neucom.2022.04.049

[10] Liu, H. (2019). Attitudes toward different types of Chinese-English code-switching. *SAGE Open*, *9*(2). https://doi.org/10.1177/2158244019853920

[11] Lyu, D.-C., Tan, T.-P., Chng, E. S., & Li, H. (2015). Mandarin-English code-switching speech corpus in Southeast Asia: SEAME. *Language Resources and Evaluation*, *49*(3), 581–600. https://doi.org/10.1007/s10579-015-9303-x

[12] Mairesse, F., Walker, M. A., Mehl, M. R., & Moore, R. K. (2007). Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research*, *30*, 457–500. https://doi.org/10.1613/jair.2349

[13] Ramírez-Esparza, N., Gosling, S. D., Benet-Martínez, V., Potter, J. P., & Pennebaker, J. W. (2006). Do bilinguals have two personalities? A special case of cultural frame switching. *Journal of Research in Personality*, *40*(2), 99–120. https://doi.org/10.1016/j.jrp.2004.09.001

[14] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter* (arXiv preprint arXiv:1910.01108).

[15] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS 2017)* (pp. 5998–6008).

[16] Warner, B., Chaffin, A., Clavié, B., Weller, O., Hallström, O., Taghadouini, S., Poli, I., & et al. (2024). Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory-efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663*. https://doi.org/10.18653/v1/2025.acl-long.127

[17] Zhang, X. (2019). Code-switching in English-Chinese ordinary conversations. *TESOL Working Paper Series*, *17*, 38–45.