Exploring Sentiment Analysis for Spanglish:

Why Sociolinguistic Context Still Matters for NLP

By

Nolan Welch

Senior Honors Thesis

Department of Romance Studies

The University of North Carolina at Chapel Hill

April 2025

Abstract

This thesis explores the role of sociolinguistic context in natural language processing (NLP), with a specific focus on sentiment analysis of Spanish-English code-switched language. Despite recent advancements in large language models (LLMs), mixed-code text remains a challenge for NLP systems, often yielding significantly lower performance compared to monolingual tasks. To investigate this gap, I evaluate the performance of both small and large multilingual language models on benchmark sentiment analysis tasks using three mixed-code datasets: LinCE, the Bangor Miami corpus, and EN-ES-CS (Aguilar et al., 2020; Deuchar, 2010; Vilares et al., 2016). I further examine whether structural measures of code-switching (CS) can serve as reliable predictors of sentiment. Finally, I conduct a time-series analysis to explore whether emotional content tends to lead or lag behind code-switch events. My results show that while the utterance-level switching metric (Gambäck & Das, 2016) is statistically associated with sentiment class, it has limited predictive value in practice, and augmenting models with this feature does not yield meaningful performance gains. Similarly, my results suggest a statistically significant leading relationship between emotionality and CS time series, though the effect size is small. These findings suggest that syntactic or frequency-based features alone are insufficient for modeling sentiment in mixed-code contexts, and that future NLP systems may benefit from integrating deeper sociolinguistic insights.

Keywords: code-switching, sentiment analysis, Spanglish, bilingualism, natural language processing, sociolinguistics, large language models

Acknowledgements

I am deeply grateful to my advisor, Dr. Lucia Binotti, for her unwavering encouragement, generous support, and insightful guidance throughout the course of this research and the writing of this thesis. Discovering fellow advocates for humanistically-centered artificial intelligence is rare, and I consider myself incredibly fortunate to have completed my thesis under the mentorship of someone with such a profound passion for the digital humanities.

Special thanks are due to the non-advising members of my honors thesis committee, Dr. Lamar Graham and Yasmin Cedamanos Del Carpio. Their invaluable feedback and sharp perspectives on these timely interdisciplinary issues enriched this project and contributed to a lively and thought-provoking defense.

I am indebted to my close friends, family, and friends who have become family over the years. Your patience and support during the long nights of writing and revision helped take this project over the finish line. Rohan, Sophie, Tanvi, David, Prithvi, Amil, Sanjay: this thesis, and my time at UNC, would not have been the same without you. Thank you.

The honors thesis support group hosted by UNC's Writing Center provided a much-needed space for accountability and discussion over the past two semesters. I am especially thankful to Maura and Dr. Gigi Taylor for their efforts in organizing this vibrant community, which reminded me that I was not alone and that I had peers to lean on for support.

Lastly, I would like to express my sincere gratitude for the institutional support I received for this project, including funding from the Alexandre Honors Carolina Fund, administered by Honors Carolina at the University of North Carolina at Chapel Hill. It is an honor to be a recipient of this award, and I hope to do it justice by continuing to pursue this research.

Framing the Problem: Linguistic Theory and NLP in 2025

William Labov, widely regarded as a founder of sociolinguistics, once remarked that he "[had] resisted the term *sociolinguistics* for years, since it implies that there can be a successful linguistic theory or practice which is not social" (1972, p. xiii). His words underscore a long-standing paradigmatic tug of war within the field of linguistics: the divide between formalist approaches that isolate language from its context, and socially grounded approaches that emphasize the embeddedness of language in culture and identity (Waugh et al., 2023).

Labov's empirically grounded, socially attuned framework, once at odds with prevailing linguistic paradigms, has since become a dominant and enduring tradition, one that highlights the complex relationships between speakers, communities, and their languages (Tagliamonte, 2015). This shift toward sociolinguistic inquiry reoriented linguistics from the formal study of abstract structures to a more holistic understanding and consideration of language as a lived, variable, and contextually rich human behavior.

Fast forward to 2025, and language is once again a central object of public fascination—though this time, not through the lens of traditional linguistics rooted in philology and the social sciences. Instead, language has become the domain of natural language processing (NLP), a subfield of computer science that has risen to cultural prominence through the widespread availability of large language models (LLMs). For many, tools like OpenAI's ChatGPT and Anthropic's Claude now represent their first real engagement with the science of language. Yet while these models are celebrated for their apparent capacity for generalized linguistic productivity, they remain statistical approximators at their core, lacking inherent capacities for reasoning, cultural interpretation, or sociolinguistic inquiry.

While the past decade's progress in NLP has been impressive, it is my view that the principles of traditional linguistic inquiry—particularly sociolinguistics—have long been undervalued in this space. In particular, the social and cultural conditions under which language is produced are often ignored by models that treat language as a neutral sequence of tokens, each semantically-equivalent combination (more or less) as valid as the next. This omission is especially salient in contexts of code-switching (CS), where language use reflects not only grammatical structure but also nuanced cues of identity and social positioning.

I intend to explore one intersection of linguistics and NLP where sociocultural insight might matter: the potential for CS patterns to augment computational sentiment analysis. Through a series of experiments, I investigate whether structural or temporal features of Spanish-English mixed-code utterances (Spanglish) carry detectable sentiment-related information, and if so, whether such information can meaningfully inform or improve NLP systems. In doing so, I aim to contribute to a broader, critical reevaluation of what it would mean to pursue sociolinguistically centered NLP, and why that might be necessary.

Argument Preview

This thesis explores multiple syntactic and structural features as predictors of sentiment in Spanish-English code-switched language (Spanglish) but finds that such approaches consistently fall short. I argue that this failure reflects a broader limitation in current NLP practices: the tendency to prioritize surface-level linguistic features while overlooking the sociolinguistic context in which language is produced. Through empirical evaluation across three lines of inquiry—language model performance, structural metrics of CS, and temporal relationships between sentiment and language switching—I demonstrate that syntactic or frequency-based features offer limited explanatory or predictive value on their own. These

findings point to the need for NLP systems that integrate deeper insights from sociolinguistics, particularly when working with non-standard or socially variable language varieties.

Motivation and Significance

The recent surge in public and academic interest in LLMs has brought NLP to the forefront of conversations about the future of language, communication, and artificial intelligence. Yet while these models have achieved remarkable feats in a variety of tasks, they continue to struggle with linguistic phenomena that fall outside of standardized, monolingual usage. I contend that sociolinguistic context—often overlooked in modern NLP research—is a key factor in improving model performance on these complex, real-world language tasks. Understanding the limitations of language models in handling mixed-code text is not only a technical challenge, but a conceptual one: it invites us to reexamine the assumptions and conventions that underpin contemporary language modeling and to reconsider how social context informs language in use.

Research Problem

Despite advances in multilingual NLP and increasing attention to linguistic diversity, current models perform poorly on mixed-code tasks such as sentiment analysis of Spanish-English utterances. This underperformance raises important questions about the limits of current modeling paradigms, especially when applied to socially embedded, non-standard varieties of language like Spanglish. I investigate whether structural or temporal patterns in CS can serve as effective signals for sentiment classification, and whether incorporating such features into NLP pipelines leads to meaningful performance gains.

Research Questions

In my investigations, I explore three interrelated research questions designed to probe the relationship between CS and speaker sentiment:

- **RQ1:** How do multilingual SLMs and LLMs perform on sentiment analysis tasks involving Spanish-English mixed-code text?
- **RQ2:** Can structural features of CS, such as the number of switch points, language ratio, or utterance-level switching metric (ULSM), predict sentiment labels?
- **RQ3:** Is there a temporal relationship between emotionally salient language and the occurrence of CS within an utterance?

Why Spanglish?

Spanglish, or Spanish-English CS, presents an especially rich site for inquiry at the intersection of sociolinguistics and NLP. As a naturally occurring and widely attested mode of communication, it reflects not only grammatical blending but also deep cultural, emotional, and contextual meaning-making (Leeman & Fuller, 2021, pp. 296–297; Lipski, 2008). The study of Spanglish benefits from access to robust datasets across modalities, including transcribed speech (e.g., the Bangor Miami corpus) and social media text (e.g., LinCE and EN-ES-CS). Furthermore, Spanglish exemplifies many of the sociolinguistic challenges faced by current NLP systems (to be discussed later), making it an ideal test case for exploring the limitations of conventional modeling approaches and the potential of sociolinguistically-informed alternatives.

Language Modeling in Context

What Is Language Modeling?

At its core, language modeling is a statistical task: it involves building a model that assigns probabilities to sequences of linguistic tokens (words, characters, subwords, and so on)

based on patterns observed in training data. For any given input, the model attempts to estimate the most likely continuation or classification, effectively learning a probability distribution over sequences in a particular language or language variety. It may be helpful to consider Grieve et al.'s (2024) conceptualization of language models: not as universal models of "language", but rather as approximators of the specific linguistic patterns present in their training data.

Though the notion of assigning probabilities to sentences was famously dismissed by Noam Chomsky, who argued that such statistical approaches fail to capture grammaticality or deep linguistic structure (1968), probabilistic models have nevertheless proven incredibly useful in modeling linguistic productivity. Unlike symbolic systems, which rely on rigid sets of rules, statistical language models can generalize from data in flexible and often surprising ways.

With that being said, it is important to acknowledge that every modeling paradigm is a choice, and all models are simplifications. NLP has largely committed to a paradigm that treats language as a sequence of tokens governed by syntactic and semantic regularities, often to the exclusion of social, cultural, and pragmatic context—arguably among the most important factors of everyday language use. I choose to interrogate that paradigm by asking: what happens when the sociolinguistic realities of language use, such as those involved in CS, are left out of the modeling process?

Small or Large?

In the current research climate surrounding NLP, I would be remiss not to mention the most dominant trend in language modeling: large language models (LLMs). Characterized by their massive parameter counts (ranging from 1 billion to over 200 billion, depending on architecture and implementation), these models demonstrate emergent capacity for long-form text generation, factual recall, and even basic forms of reasoning and pragmatic inference (Li et

al., 2024). Some studies also point to their apparent capacity to simulate register variation and adapt stylistically to different social personas, raising questions about the extent to which these models encode or reproduce sociolinguistic phenomena (Deshpande et al., 2023; Huang et al., 2024; Tseng et al., 2024; Yang et al., 2024).

Wang et al. (2024) note that the definition of a "large" language model is far from standardized, with criteria including emergent behavior, domain-specific competencies, and simple parameter count. For clarity, we will refer to models with 1 billion parameters or fewer as small language models (SLMs), and all others as LLMs.

Despite their dominance, LLMs are not unambiguously superior across all NLP tasks. In sentiment analysis, for instance, Zhang et al. (2023) find that LLMs outperform SLMs, though only marginally, suggesting that traditional feature engineering and smaller, targeted models may still offer strong performance in this domain. This is particularly relevant when modeling phenomena like CS, where sociolinguistic nuance may not be adequately captured by massive general-purpose language models alone.

As such, this paper positions itself within a growing body of work that revisits traditional linguistic theory, especially sociolinguistics, as a source of insight for designing and evaluating language models. Even as LLMs continue to grow in scale and influence, there remains significant value in asking what kinds of language patterns are still under-modeled, and how we might better capture them—perhaps not through scale, but through sociolinguistic awareness.

Background and Related Work

Contemporary Sentiment Analysis

What is Sentiment Analysis?

Sentiment analysis is a core NLP task that consists of determining a base emotional state (e.g., positive, negative, or neutral) given a text document (Jiawa et al., 2021; Jin et al., 2023; Wankhade et al., 2022). This task has wide applications, from corporate market research to social attitude examination; furthermore, it is a basic component of automated systems for parsing and generating meaningful representations of human-generated language.

How It's Typically Modeled

Zhang et al. (2023) highlight three broad categories of sentiment analysis tasks: sentiment classification (SC), aspect-based sentiment analysis (ABSA), and multifaceted analysis of subjective texts (MAST). We will primarily discuss the former.

In sentiment classification, sentiment analysis is modeled as a sequence labelling task. Given a sentence (or, equivalently, a string of tokens or sub-words), a model predicts the appropriate sentiment label. This label might be binary (positive or negative), ternary (positive, negative, or neutral), or follow a more fine-grained scale (e.g., positivity rating from 1 to 5).

Sentiment classification may be applied at different levels of granularity. At the document level, models assess the overall sentiment of an entire text. At the sentence level, they evaluate the emotional tone of individual sentences. Finally, aspect-based sentiment analysis focuses on identifying sentiment toward specific components or features mentioned in the text (e.g., an online review for a product might have different sentiments towards its color and build quality).

Modern sentiment analysis models often rely on transformer-based architectures pre-trained on large corpora and fine-tuned for the sentiment classification task (Jin et al., 2023;

Vaswani et al., 2023; Wankhade et al., 2022; Zhang et al., 2023). While these models achieve state-of-the-art results on monolingual datasets, their performance tends to suffer in mixed-code contexts, where syntactic irregularities and sociocultural nuance play a more central role.

Mixed-Code Language and NLP

Given the global prevalence of multilingualism, it is unsurprising that NLP tasks have been applied to mixed-code data in an effort to improve computational models of this widespread mode of communication (Chatterjere et al., 2020). Specialized models that account for the peculiarities of CS have been developed for natural language generation, natural language understanding, sentiment analysis, and other NLP tasks (Aryal et al., 2022; Doğruöz et al., 2021; Nazir et al., 2025; Srinivasan & Subalalitha, 2023; Winata et al., 2023).

Despite significant advancements in NLP and extremely high accuracy (often exceeding 90 percent) in monolingual sentiment analysis tasks (*IMDb Benchmark (Sentiment Analysis)*, 2025), sentiment analysis for mixed-code contexts remains a persistent challenge; the best-performing models for mixed-code tasks typically only reach between 58 and 62 percent classification accuracy (*LinCE Leaderboard*, 2025). I investigate the causes of this disparity in performance and explore whether it arises from a lack of training data for underresourced languages or reflects deeper limitations in current modeling approaches. In other words: can mixed-code language modeling be improved using the same strategies that work for monolingual data, or do we need to develop fundamentally novel methods for this class of linguistic behavior?

In NLP, improving model performance on underresourced languages typically involves supplementing training data in the target language (Gibadullin et al., 2019; Hedderich et al., 2021; Mabokela et al., 2023; Ranathunga et al., 2021). For sentiment analysis, Zeng (2024) shows that supplying synthetic code-switched data generated by LLMs improves classification

accuracy, offering a promising direction for model augmentation. Additionally, Volkova et al. (2013) demonstrate the limited but present predictive power of gender-based differences in the production of mixed-code writing for sentiment classification.

Despite the availability of numerous Twitter datasets for NLP and sentiment analysis, including several with mixed-code Spanish-English utterances (Aguilar et al., 2020; Mathur & Shrivastava, 2024; Pérez et al., 2022, 2024; Sutar et al., 2023; Vilares et al., 2015, 2016), I argue that relying solely on text-based data to model native speaker CS is insufficient. CS, as a sociolinguistic phenomenon, is best understood as arising from spontaneous, real-time interactions between speakers using two or more languages in conversation (Cedden et al., 2024; Gardner-Chloros, 2009). The key term here is *spontaneous*: the moment-to-moment decisions that multilingual speakers make in live discourse give rise to CS's variability and distinct character. Unlike monolingual modes of communication, which follow relatively stable grammatical conventions, CS lacks rigid structure by its very nature.

Because spontaneous CS is primarily an oral phenomenon, I posit that written communication, which tends to be a relatively more planned form of discourse under Ochs' (1979) dichotomy, fails to capture the core features of CS, and is therefore less representative of the linguistic behaviors I aim to study. For this reason, my work focuses on spoken mixed-code conversations, particularly from the Bangor Miami corpus, which includes 56 manually annotated recordings with detailed language identification (Deuchar, 2010). By analyzing sentiment in transcripts of spoken language rather than social media posts, I aim to give a fuller account of the spontaneous nature of CS and its relationship with speaker emotion.

Code-Switching and Emotion

The relationship between CS and the emotional speaker state has been previously explored from various theoretical and experimental frames of analysis. Dewaele (2010) finds an inverse correlation between the recency with which a language was learned (e.g., L2 is acquired later than L1) and the likelihood that it will be used to express emotion. In other words, the earlier a language is learned, the more likely it is to be used for emotionally expressive content. This observation has implications for sentiment analysis in mixed-code contexts, suggesting that incorporating information about a speaker's native or dominant language could improve classification accuracy.

Additionally, Pavlenko (2008) highlights the importance of emotionality and emotion-laden vocabulary in multilingual CS. She argues that the type and strength of emotional state may influence not only language choice, but also the timing and function of CS events. These insights play a central role in shaping multilingual discourse, and thus should be considered in computational models seeking to interpret sentiment in code-switched language.

Sentiment Analysis in Mixed-Code Contexts

Prior work (Hovy, 2015; Volkova et al., 2013) supports the idea that performance across a variety of NLP tasks can be improved by conditioning on sociocultural factors. This suggests that patterns of language use are influenced to a measurable and informative degree by extralinguistic factors, a notion for which the field of sociolinguistics provides strong evidence. The key question, then, is why those sociocultural factors are left on the table, so to speak, when creating computational models of language. I will discuss this in greater detail later on.

Although sentiment analysis has benefited from advances in deep learning and the advent of LLMs, most approaches still focus heavily on surface-level lexical and syntactic features. In

mixed-code contexts, however, language choice itself may be a crucial signal: CS often reflects shifts in speaker stance, audience orientation, and emotional salience, as highlighted by Pavlenko (2008). Ignoring such signals risks flattening the rich sociolinguistic landscape that informs meaning in code-switched bilingual discourse.

Why Sociocultural Factors Matter

On a fundamental level, the study of language modeling, and by extension NLP and its associated subtasks, cannot be disentangled from the study of language as a sociocultural phenomenon. To treat language as a mere set of lexically-constrained contextual probability distributions is to discount the rich and varied sociocultural dimensions of language as it is lived and performed in human communities.

Contemporary NLP methods are undeniably impressive in their scalability and in their ability to produce language that is often fluent, semantically coherent, and pragmatically appropriate. Still, they remain probabilistic models that exhibit only shallow syntactic, semantic, and pragmatic awareness. Despite their apparent linguistic fluency, these systems often fail to engage with the deeper structures and social functions that define language in use.

Because LLMs' output is tightly coupled to the data they are trained on, performance disparities across languages are to be expected. Languages that are underrepresented in datasets like Common Crawl often see degraded performance on NLP tasks that require advanced understanding of phrase structure, cultural pragmatics, and other usage-based phenomena. In these cases, the gap between statistical language modeling and lived linguistic experience becomes most visible.

A key question I aim to raise is whether traditional approaches to language modeling are reliable when applied to non-standard or minoritized linguistic varieties. For example,

Spanish-English code-mixed text is widespread online, yet LLMs consistently underperform on NLP tasks in mixed-code contexts compared to their performance on monolingual benchmarks. Why is this? Is there simply insufficient training data to support broad generalization in these contexts? Or might it be that CS taps into dimensions of language—social identity, community norms, interspeaker variation—that go beyond syntax and semantics? Perhaps sociocultural factors are not external to language modeling, but rather central to it.

Methodology

Research Design Overview

To investigate the persistent underperformance of language models on sentiment analysis tasks involving mixed-code Spanish-English data, this study adopts a multi-pronged approach that combines model benchmarking, feature-based statistical analysis, and time-series experimentation. This design reflects the multifaceted nature of the problem: while prior work indicates that NLP models struggle with code-switched data, the reasons for this remain underexplored (Winata et al., 2023). Rather than assuming a single cause, such as a lack of training data, this study treats the issue as potentially arising from multiple linguistic and statistical factors.

Accordingly, I pursue three interrelated lines of inquiry, each designed to probe a distinct aspect of the problem:

1. **RQ1:** *How do multilingual small and LLMs perform on mixed-code sentiment analysis tasks?* This question establishes a performance baseline by evaluating models of varying scale across two datasets (LinCE and EN-ES-CS), allowing for direct comparison between architectures and prompting strategies.

2. **RQ2:** *Can utterance-level structural features of CS serve as predictors of sentiment class?* Here, I explore whether simple syntactic indicators, such as language ratios and CS frequency, convey useful sentiment-related information.
3. **RQ3:** *Is there a temporal relationship between emotionally salient language and code-switch events within an utterance?* Finally, I use time-series cross-correlation techniques to test whether emotional content precedes or follows CS, hoping to reveal latent discourse-level patterns, if any exist.

Together, these threads offer a holistic, layered view of where current approaches to mixed-code sentiment analysis succeed, where they fall short, and what types of information might help improve future computational methods.

Datasets

For the purpose of my analyses, I examine three datasets: the Bangor Miami corpus, the LinCE dataset, and the EN-ES-CS dataset. In the sections below, I provide a brief description of each dataset, as well as my parsing and preprocessing methods.

Bangor Miami Corpus

The Bangor Miami corpus (Deuchar, 2010) consists of transcribed oral conversations between a study participant, dubbed “María”, and other speakers from Miami, Florida. The transcripts have been manually annotated with detailed language identification (LID) features, including markers to indicate whether morphological features are mixed from English and Spanish. There is also part-of-speech (POS) tagging, as well as translations to English for all Spanish sentences. The corpus is publicly available as a part of the BilingBank database (a component of TalkBank) and specified according to the CHAT transcription format (MacWhinney, 2000, 2019, 2020).

It is worth noting that a large portion of the code-switched speech in this corpus is sourced from a single speaker, and thus may not be representative of broader sociolinguistic patterns. However, the corpus' detailed linguistic annotations and parallel English translations make it uniquely useful for exploratory analyses of CS dynamics.

LinCE Dataset

The LinCE dataset (Aguilar et al., 2020) is a multilingual, multi-task dataset intended to serve as an evaluator of language model performance in mixed-code contexts. The dataset includes Spanish-English tweets labeled for a range of NLP tasks, including part-of-speech tagging, language identification, and sentiment analysis. The data is provided split into a training set (12,194 samples), a development set (1,859 samples), and a test set (for evaluation on the online LinCE leaderboard), and is specified in CoNLL format. For the purposes of this study, I focused exclusively on the sentiment classification subset of the dataset. While less richly annotated than the Bangor Miami corpus, LinCE offers a more diverse and representative sample of mixed-code language in a written social media context.

EN-ES-CS Dataset

The EN-ES-CS dataset (Vilares et al., 2016) consists of Spanish-English code-switched tweets annotated for sentiment. It includes 3,062 tweets unevenly distributed across three sentiment categories: positive (963 samples), negative (786 samples), and neutral (1,313 samples). Compared to LinCE, the EN-ES-CS dataset is smaller, but provides good variety for controlled sentiment classification experiments. I selected this dataset for its relative structural simplicity and because it has been used as a benchmark in several prior studies of sentiment analysis in mixed-code contexts (Ahuja et al., 2023; Barbieri et al., 2020; Jose et al., 2020; Mabokela et al., 2023).

Preprocessing Pipeline

For all datasets, I applied a consistent preprocessing pipeline to ensure comparability and to prepare the data for analysis and model training. The pipeline includes the following steps:

1. Lowercasing all tokens
2. Removing URLs, hashtags, mentions, and non-ASCII characters
3. Lemmatization using language-specific models from spaCy (Honnibal et al., 2020)
4. Stopword removal (filler words or words with low emotional content)
5. Punctuation scrubbing

As an example, the unprocessed input string "@_OmarReyes jajajajj me avisas, que lo necesito :) :) :) saludos" is normalized and cleaned as "jajajajj avisa necesitar saludos".

For structural feature extraction, including the number of code-switch points and the Utterance-Level Switching Metric (ULSM), I used language identification tags when available (e.g., in the Bangor Miami corpus) or heuristic-based labeling based on token-level language detection using the *Lingua* Python library (Stahl, 2021/2025). The preprocessed text and associated features were then formatted into comma-separated values (CSV) files for downstream model training and evaluation.

RQ1: Benchmarking Language Model Performance

This experiment evaluates the performance of small and large multilingual language models on sentiment analysis tasks in Spanish-English mixed-code contexts. The objective is to establish a performance baseline and identify whether scale, fine-tuning, or prompting strategies significantly influence classification outcomes in these settings.

Model Selection

Two categories of models were tested. See Appendix for additional model details.

- **Small Language Models (SLMs):**
    - LM1 (BERT family, 109M parameters)
    - LM2 (DistilBERT family, 135M parameters)
    - LM3 (RoBERTa family, 279M parameters)
- **Large Language Models (LLMs):**
    - LM4 (GPT family, estimated 8B parameters)
    - LM5 (GPT family, estimated 200B parameters)

Evaluation Protocol

For this experiment, I evaluated models on the EN-ES-CS and LinCE datasets, as these text-based datasets lend themselves easily to multilingual and code-switching analysis.

As prior work has shown, language model performance on most mixed-code NLP tasks is quite poor, especially when compared to LM performance on the same tasks in monolingual contexts (Sitaram et al., 2020; Winata et al., 2023). In order to investigate this phenomenon and explore its boundaries—for example, whether larger model sizes correlate with improved task accuracy—I evaluated five language models on the LinCE and EN-ES-CS datasets.

SLMs were evaluated directly on both datasets using pretrained sentiment classification heads. LLMs were evaluated only on the LinCE dataset using three prompting strategies: zero-shot prompting, few-shot prompting, and chain-of-thought prompting. Prior work motivates the use of these prompting strategies as a means of improving LLM performance; see Schulhoff et al. (2025) for a thorough review of current research on prompt engineering. For the full prompts used in these LLM experiments, see Appendix. A brief definition of each strategy is given below.

- **Zero-shot prompting:** The LLM is only given the classification task, with no examples.

- **Few-shot prompting:** The LLM is given the classification task, plus 3 examples of desired classification behavior.
- **Chain-of-thought prompting:** The LLM is encouraged to provide its classification after "thinking out loud" in a step-by-step manner.

Prior work, including Schulhoff et al. (2025), suggests that incorporating information regarding the task domain at prompt time can improve LLM performance. To further break down the factors involved in language model performance for the mixed-code sentiment classification task, I performed each LLM experiment in two settings:

- **CS-naive:** No explicit information regarding CS was provided to the model during prompting.
- **CS-aware:** In-context learning (ICL) examples included code-switched inputs.

Each LLM was run across three trials per prompting condition. The primary metrics reported are accuracy, precision, recall, and F1 score (weighted average).

RQ2: Structural Code-Switching Features as Predictors of Sentiment

In this portion of the study, I investigated whether simple structural features of code switching in an utterance could serve as predictive features for sentiment classification. I focused on three utterance-level features in particular: the number of code-switch points, the proportion of the sentence that was in English or Spanish, and the utterance-level switching metric (ULSM), proposed by Gambäck and Das (2016) (see Appendix). The goal of this experiment was to evaluate whether these features alone, without their lexical context, could meaningfully predict sentiment labels in mixed-code utterances.

Feature Engineering and Augmentation

To begin, I augmented the EN-ES-CS dataset (Vilares et al., 2016) by extracting the following structural features on a per-utterance basis:

- **Number of code-switch points:** The number of times the language changes between adjacent words.
- **Language ratio:** The proportion of words in the sentence that are in English, as a proxy for the bilinguality of the utterance.
- **Utterance-Level Switching Metric (ULSM):** A normalized score that quantifies the amount of CS relative to utterance length. ULSM was computed using the majority language in the utterance as the matrix language; see Appendix for more.

These features were calculated using custom scripts that leveraged token-level language labels, which were generated heuristically for this dataset using existing tools and dictionaries.

Preprocessing and Normalization

Prior to feature extraction, I applied several preprocessing steps to ensure consistency across utterances. These included lemmatization, removal of stopwords and punctuation, and normalization to lowercase. The resulting cleaned text was then used both to compute structural features and to train benchmark classifiers. For numerical features, I standardized values to facilitate statistical comparison and modeling.

Statistical Analysis

To assess whether any of the three structural features were significantly associated with sentiment class, I performed a Kruskal-Wallis H-test, a non-parametric alternative to ANOVA suitable for ordinal and non-normally distributed data (1952). Because the sentiment label is a

three-level categorical variable (positive, negative, neutral), this test was applied separately for each structural feature, with $\alpha = 0.05$ and two degrees of freedom.

Post Hoc Testing

For features that showed a statistically significant result in the Kruskal-Wallis test, I conducted a Dunn's test with Bonferroni correction (1961) to identify which sentiment category pairs differed significantly. This enabled finer-grained analysis of whether CS metrics differed significantly between specific sentiment classes (e.g., negative vs. neutral).

Baseline Classifier Using ULSM

To test the practical utility of the strongest-performing structural feature, ULSM, I trained a naive logistic regression classifier to predict sentiment class using ULSM as the sole feature. To improve interpretability, I collapsed sentiment into a binary task: classifying utterances as either negative or non-negative. Model performance was evaluated using weighted F1 score and area under the ROC curve (AUC) (Li, 2024).

Feature-Enhanced Model Evaluation

Finally, I tested whether adding ULSM as an additional feature would improve the performance of traditional text-based classifiers. Using the same preprocessing pipeline and training data, I trained and evaluated five standard machine learning models—dummy classifier (most frequent class prediction), decision tree, k-nearest neighbors, naive Bayes, and random forest—under two conditions: text-only baseline, and text augmented with the ULSM feature. Text features were given as text frequency-inverse document frequency (tf-idf) values calculated with the scikit-learn Python module (Pedregosa et al., 2011). This allowed for a controlled comparison of model performance with and without CS information.

RQ3: Temporal Alignment Between Emotion and Code-Switching

In this final line of inquiry, I conduct a time-series cross-correlation analysis to determine whether emotionally salient language (as measured by token-level emotional arousal and valence) serves as a leading or lagging indicator of code-switch points within an utterance. This experiment aims to probe the discourse-level dynamics of sentiment and language choice.

In my approach, I model a sentence *S* as a sequence of temporally-ordered word-level observations, each of which is associated with some sentimental content and a language. This enables the application of time-series based analysis techniques, treating each word as a separate time step. For my analysis, I focused on two signals: language switching and speaker sentiment.

Language Switching Signal

I defined the language-switching signal as a time series of length *n*, where *n* is the number of words in the sentence. At each time step *t*, this signal takes on a value according to a binary function CS(t) depending on whether a code-switch (change in language) has occurred relative to the word at time step *t-1*:

- CS(t) = 0 if t = 0
- CS(t) = 0 if Lang(t) = Lang(t-1)
- CS(t) = 1 if Lang(t) ≠ Lang(t-1)

To illustrate this labeling method, several examples of mixed-code sentences from the EN-ES-CS corpus (Vilares et al., 2016) and their CS signals are given below.

Sentence	Code-Switching Signal
My brother was like next time <u>que</u> venga Te Lo Presento	[0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0]
Now <u>Eso</u> Si Fue boda del Año	[0, 1, 0, 0, 0, 0, 0]
I wanted to dance <u>Reggeaton</u> Pero no Tenia	[0, 0, 0, 0, 1, 0, 0, 0, 1, 0]

dancing partner	

For the word-level language identification Lang(t), I used the hand-labeled language annotations provided in the Bangor Miami corpus. The corpus' annotations identify multiple language types beyond just English and Spanish: there are unique tags for mixed-language words, terms of ambiguous linguistic origin, and words from neither English nor Spanish (e.g., French words or non-words). For simplicity's sake, I adopted a naive approach, treating each unique tag as a unique language for the purpose of calculating CS points.

Sentiment Signals

For this investigation, sentiment signals are a pair of time series of length *n*, where *n* is the number of words in the sentence. For each word, I used a multilingual RoBERTa model fine-tuned for sentiment analysis (LM3; see Appendix) to calculate valence and arousal values, drawing on Russell's (1980) two-dimensional "circumplex" model of emotion. We approximated these values using LM3 as follows:

- **Valence:** Approximated as (Probability of positive class) – (Probability of negative class). A higher value means the term is more likely to be positive.
- **Arousal:** Approximated as the maximum probability between P(positive) and P(negative), where "P(X)" means "the probability that sentiment class is X". A higher value means the term is more strongly emotionally biased.

Results

RQ1: Model Performance

Performance of Small Language Models

Experimental results are presented below. Best scores are underlined for each dataset.

Model	Dataset	Accuracy	Precision	Recall	F1
LM1	EN-ES-CS	0.63	0.65	0.63	0.62
	LinCE	0.46	0.58	0.46	0.46
LM2	EN-ES-CS	0.43	0.51	0.44	0.37
	LinCE	0.53	0.48	0.53	0.47
LM3	EN-ES-CS	0.66	0.66	0.66	0.66
	LinCE	0.52	0.58	0.52	0.53

Performance of Large Language Models

The figures provided below are F1 scores averaged across three trials. The top F1 score across models, prompting strategies, and awareness settings is underlined.

Model	Prompting method	CS-Naive	CS-Aware
LM4	Zero-shot	0.557	0.537
	Few-shot	0.543	0.550
	Chain-of-thought	0.550	0.537
LM5	Zero-shot	0.453	0.513
	Few-shot	0.447	0.400
	Chain-of-thought	0.510	0.520

RQ2: Structural Features and Sentiment

Kruskal-Wallis and Dunn's Post Hoc Results

Of the three features tested, only the ULSM displayed a statistically significant relationship with sentiment category. The Kruskal-Wallis test yielded $H = 16.76$, $p = 0.0002$, surpassing the critical chi-squared value of 5.99 for $\alpha = 0.05$ with two degrees of freedom. In contrast, the number of code-switch points ($H = 1.98$, $p = 0.372$) and the English word ratio ($H =$

0.61, p = 0.737) showed no significant association with sentiment class. (Note that, since we only model English and Spanish, we do not need to model English ratio and Spanish ratio separately; Ratio(Spanish) = 1 – Ratio(English), so any relationship with one would hold for the other.).

A follow-up Dunn's test with Bonferroni correction revealed that the difference in ULSM was statistically significant for comparisons between negative and neutral sentiment (p = 0.0002), as well as between negative and positive sentiment (p = 0.004), but not between positive and neutral categories.

Naive Logistic Regression Using ULSM

To further assess ULSM's practical utility, I trained a simple logistic regression model to classify sentiment as negative or non-negative based solely on the ULSM value. The model achieved a weighted F1 score of 0.57 and an area under the ROC curve (AUC) of 0.54, suggesting that while the metric carries some signal, it is a poor standalone predictor of negative vs. non-negative sentiment class. See Appendix for the ROC curve.

Text + ULSM Model Performance

Model	Baseline F1 Score	Augmented F1 Score
Dummy classifier	0.40	0.40
Decision tree	<u>0.47</u>	0.45
K nearest neighbors	0.49	0.49
Naive Bayes	<u>0.51</u>	0.49
Random forest	0.46	0.46

The results of this experiment indicated that in all but one case, the inclusion of ULSM did not improve model performance, and in some cases slightly reduced it. For instance, the

decision tree classifier's F1 score dropped from 0.47 to 0.45 when ULSM was added, and naive Bayes dropped from 0.51 to 0.49. These findings indicate that while ULSM may show statistically significant differences across sentiment categories, it lacks sufficient predictive strength to improve classifier accuracy in practice.

RQ3: Cross-Correlation Results

Observed Patterns

In the Bangor Miami dataset, the peak correlation between emotional arousal and CS events occurred at a lag of +1, indicating that emotional content tends to precede code-switch events by one token. This alignment supports the hypothesis that CS may be responsive to emotionally charged content, rather than anticipatory.

Significance Testing

To evaluate whether these observed correlations were meaningful or the result of random noise, I constructed a null distribution by randomly shuffling sentiment values within each utterance and recalculating peak correlations for 10,000 such permutations. The empirical p-values, based on the proportion of shuffled trials with greater or equal peak correlation, were effectively zero for both datasets. Additionally, I calculated z-scores based on the null distribution mean and standard deviation. The observed peak correlation was 0.0149, and the null distribution's mean and standard deviation were $\mu = 0.0015$ and $\sigma = 0.00084$. This puts the peak observation at $Z = 15.8$, $p < 2.4 * 10^{-55}$ (assuming that the null distribution is roughly normal). These results strongly reject the null hypothesis that the observed alignment between sentiment and CS is random. The alignment is small in magnitude, and further experimentation should aim to confirm this phenomenon across multiple datasets and modalities (e.g., other datasets and text-based datasets).

Discussion

Interpreting RQ1: Benchmarks and Implications

SLMs Underperform on Mixed-Code Sentiment Tasks

Despite strong performance on monolingual benchmarks such as IMDb or SST-2 (Maas et al., 2011; Socher et al., 2013), none of the SLMs achieved F1 scores above 0.66 on mixed-code datasets. Notably, even models explicitly trained for multilingual sentiment analysis performed only moderately well, particularly on the LinCE dataset. This suggests that CS presents structural challenges that are not addressed by standard pretraining strategies.

LLMs Show Inconsistent Improvements Despite Scale

The performance of LMs 4 and 5 was mixed. LM4, a smaller model, often outperformed the much larger LM5—a surprising result that suggests diminishing returns from scale in mixed-code contexts when fine-tuning or domain-specific adaptation is absent. Across all prompting modes, the highest F1 score for any LLM was 0.56, which is underwhelming for models that often exceed classification accuracies of 0.90 on monolingual tasks (*IMDb Benchmark (Sentiment Analysis)*, 2025).

In-Context Priming With CS-Aware Examples Yields Marginal Gains

Priming LLMs with code-switched examples (CS-aware setting) produced only minor gains in most cases. The only reliably positive shift was observed in the zero-shot setting for LM5, where CS-aware priming led to an F1 increase of 0.06. However, in few-shot scenarios, CS-aware performance even declined slightly, possibly due to example clutter or increased confusion around language boundaries.

Prompting Strategy Mattered Less Than Expected

Chain-of-thought prompting, often beneficial for logical reasoning tasks, did not substantially outperform zero-shot or few-shot prompting. In fact, differences between prompting styles were minimal across all tested configurations, suggesting that sentiment classification on mixed-code data may not benefit from reasoning-based prompting in its current form.

Interpreting RQ2: Code-Switching as a Structural Signal

ULSM Is Statistically Significant but Weak in Practice

The ULSM was the only CS feature to show a significant difference across sentiment categories. However, the post hoc analysis suggests that this signal is primarily driven by contrast between negative sentiment and other classes, with positive vs. neutral remaining indistinguishable.

Poor Performance of Naive Classifier Confirms Weak Predictive Value

Although statistically significant, the ULSM is not a strong predictor of sentiment in practice. An AUROC of 0.54 and an F1 score of 0.57 are marginally better than random chance for the binary classification task (negative vs. non-negative).

Augmenting Text-Based Models With ULSM Does Not Improve Performance

Across all model architectures tested, the inclusion of ULSM as a feature failed to improve or even maintain performance relative to the baseline. In some cases (e.g., Decision Tree, Naive Bayes), performance slightly decreased.

Structural Code-Switching Metrics May Lack Theoretical Power

While it's encouraging that ULSM captured some signal, the fact that neither switch count nor language ratio were predictive suggests that purely syntactic or frequency-based

metrics may not encode the emotional nuance present in mixed-code communication. This supports the idea that future work should look to sociolinguistic or discourse-level features.

Interpreting RQ3: Emotion Leading Code-Switching

Summary of Findings

The results from the cross-correlation analysis support the idea that emotional salience precedes CS events. The consistent peak correlation at a +1 word lag in the Bangor Miami corpus indicates that emotionally intense words tend to come just before a language switch, suggesting a reactive rather than anticipatory function of CS in emotional moments.

Effect Size and Practical Significance

While the results are statistically significant, the small magnitude of the observed correlational effect ($r = 0.0149$) means they are of limited practical utility. Although these results are unlikely to have occurred by chance, this weak correlation suggests limited predictive power in real-world applications, especially when relying on simple, token-level sentiment estimations.

Modeling Implications

These findings contribute to the growing body of evidence that linguistic features like CS are influenced by more than grammatical constraints, and also respond to social and emotional context. As such, computational models that aim to predict or understand CS behavior may benefit from integrating pragmatic and sociocultural cues (such as speaker identity, topic of conversation, or setting) alongside lexical and morphosyntactic cues.

Limitations and Future Directions

This analysis is limited by its reliance on estimated sentiment scores and a limited word-level resolution. Future research could explore more robust emotion modeling techniques, incorporate discourse-level or prosodic features, and test across a wider range of bilingual

datasets to validate the generality of the finding. Additionally, future work could improve on this experiment's naive approximations of valence and arousal by incorporating more advanced models that score sentences on their emotional valence and arousal, such as VADER (Hutto & Gilbert, 2014) or SentiStrength (Thelwall et al., 2012).

Methodological Reflection

With the benefits of hindsight and a deeper understanding of the datasets and modeling challenges involved, I would like to reflect on several limitations of my experimental methodology and highlight areas where future research could improve upon this work.

Prioritizing Oral Language Data

One key insight is that CS, as a language contact phenomenon largely defined by spontaneous, context-dependent decision making, is best studied in oral, transcribed conversation. Unlike written or semi-planned communication (e.g., social media posts), natural speech more reliably captures the real-time cues that influence CS behavior. In this light, an ideal experimental setup would have focused more heavily on corpora like the Bangor Miami corpus or similar transcribed interviews with native Spanish-English bilinguals. Even more effectively, future work might involve designing sociolinguistic interview protocols that intentionally elicit emotionally charged speech, thereby improving the validity and accuracy of analyses for sentiment analysis tasks.

Interpreting Small Effect Sizes

Although this study finds statistically significant evidence of a relationship between CS behavior and speaker sentiment, the effect sizes are small and unlikely to support practical applications in their current state. Indeed, it is possible that these weak associations are a result of noise introduced by coarse-grained measures of sentiment and structural CS. Future work

aiming to connect sociolinguistic patterns with computational predictions should seek phenomena with stronger empirical precedents and consider higher-resolution emotion modeling strategies.

Limitations in Time-Series Design

The design of the time-series analysis used to examine temporal correlations between sentiment and CS could also be improved. Ideally, sentiment analysis would have been performed on monolingual data, such as the English translations found in the Bangor Miami corpus. Monolingual inputs tend to produce more accurate and consistent outputs from current language models. However, this approach was ultimately infeasible due to alignment issues between the English translations and the original mixed-code sentences; since they differ in structure and length, the required isomorphic sequences for cross-correlation were not available. Alternative alignment techniques or attention-based models may prove helpful in future research.

Exploring Finer-Grained Sentiment Tasks

This project focused primarily on utterance-level sentiment classification, a relatively coarse task. While appropriate for a first step, future work may benefit from exploring more fine-grained sentiment subtasks, such as aspect-based sentiment analysis. These more nuanced tasks might reveal patterns missed by global labeling approaches and would be better suited to capturing the rich interplay between CS and emotion.

Toward More Nuanced Experimental Design

A more refined experimental design, with greater attention to exploratory analysis in the early stages and more careful selection of syntactic and structural features of CS, may yield deeper insights into the connection between bilingual CS and emotional expression. As computational tools become more sophisticated, aligning their outputs with human

sociolinguistic intuition will become increasingly essential for advancing NLP in diverse, multilingual contexts.

Reframing the Problem: Sociolinguistic Awareness in NLP

Language is nothing if not social. As Labov argued, it is both foolhardy and nearly impossible to pursue a study of linguistics that is not firmly rooted in an understanding of the social context in which it occurs. While contemporary language models are trained on large-scale text corpora consisting of (primarily) authentic examples of language in use, the social dimension of language is all but lost, hidden behind loose associations with factors unknown to the LM. Humans incorporate sociocultural information on a regular basis when parsing and producing language, with profound pragmatic implications for how an utterance is interpreted.

Consider, for instance, the sentence “What’s the problem, mija?” spoken in two different contexts: one by a Hispanic bilingual parent in a moment of tenderness, and the other by a monolingual outsider in a sarcastic tone. Without sociocultural grounding, a model cannot distinguish the affectionate use from the potentially mocking one. This limitation underscores a broader point: many of the most meaningful distinctions in language use hinge not on word order or token frequency, but on speaker identity, intention, and social relationship.

Hopefully, the reader is now convinced of the importance, if not the absolute necessity, of incorporating sociocultural awareness into computational language models in conjunction with their base awareness of token relations and distribution. But this leads to a crucial question: How can we build models that account for human-centered context? Clearly, demographic information is not always available in NLP tasks, and in fact rarely is; much of the discourse used for NLP tasks is captured from semi-anonymous public-facing platforms such as Reddit, IMDb, Amazon, and Twitter/X, where traditional demographic signals are sparse, noisy, or absent entirely.

A core question arises: Can a more complete model of this extralinguistic and sociocultural information be attained by feeding more data into existing language models, or must we develop novel approaches that represent these factors as first-class components of linguistic meaning? Tackling this challenge may mean breaking with the current paradigm of inflated model size in favor of targeted architectures that prioritize sociolinguistic variables. One possibility is the integration of metadata-rich corpora, including information such as speaker background information or community-oriented language patterns, into training pipelines. Another is to adapt NLP evaluation frameworks to measure performance on context-sensitive, socially embedded tasks.

Incorporating demographic characteristics such as age, gender identity, regional background, and socioeconomic status into linguistic models could significantly improve their ability to capture the nuance of language use across social groups. The utility of this approach, as I see it, would be twofold: first, it would enhance predictive accuracy in socially complex modeling tasks, such as dialect identification, politeness detection, or code-switching prediction. Second, it would enable models to represent the full diversity of human language in a faithful manner, taking into consideration the social factors that lead to this embodied linguistic practice in the first place. In the long term, these developments could lead to a new generation of NLP tools that are not only more precise but also more socially aware, improving both theoretical models of language and practical applications in multilingual and multicultural contexts.

Conclusion

I have explored a specific, illustrative challenge at the intersection of sociolinguistics and computational language modeling: the interpretation of sentiment in Spanish-English mixed-code language. Through empirical investigations of structural and temporal features in

code-switched data, I have shown that while certain linguistic patterns (including ULSM and sentiment-CS alignment) may correlate with speaker emotion, their predictive utility remains limited. These findings suggest that the shortcomings of current NLP approaches in mixed-code contexts stem not merely from insufficient data or model size, but from a more fundamental oversight: the neglect of sociocultural nuance.

As LLMs continue to shape public perceptions of what language technology can and cannot do, we must critically assess the assumptions embedded within them. Language is not just lexical semantics glued together by a uniform syntax to form meaning—it is also identity, emotion, and context. If we hope to build NLP systems that truly understand and serve the richness of human communication, we must move beyond token-level correlations and toward models that acknowledge language as a socially embedded act.

This work aims to be a small step toward that broader vision. By drawing from the traditions of sociolinguistic inquiry and applying them rigorously to contemporary NLP tasks, we can begin to ask deeper, more consequential questions about language, power, and representation. Only by bridging these fields can we design technologies that are not merely functional, but socially attuned—technologies that better capture the vibrant, complex realities of human communication and expression.

References

Aguilar, G., Kar, S., & Solorio, T. (2020). LinCE: A centralized benchmark for linguistic code-switching evaluation. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 1803–1813). European Language Resources Association. https://aclanthology.org/2020.lrec-1.223

Ahuja, K., Diddee, H., Hada, R., Ochieng, M., Ramesh, K., Jain, P., Nambi, A., Ganu, T., Segal, S., Axmed, M., Bali, K., & Sitaram, S. (2023). *MEGA: Multilingual Evaluation of Generative AI* (No. arXiv:2303.12528). arXiv. https://doi.org/10.48550/arXiv.2303.12528

Aryal, S. K., Prioleau, H., & Washington, G. (2022). *Sentiment classification of code-switched text using pre-trained multilingual embeddings and segmentation* (No. arXiv:2210.16461). arXiv. https://doi.org/10.48550/arXiv.2210.16461

Barbieri, F., Camacho-Collados, J., Neves, L., & Espinosa-Anke, L. (2020). *TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification* (No. arXiv:2010.12421). arXiv. https://doi.org/10.48550/arXiv.2010.12421

Camacho-Collados, J., Rezaee, K., Riahi, T., Ushio, A., Loureiro, D., Antypas, D., Boisson, J., Espinosa-Anke, L., Liu, F., Martínez-Cámara, E., Medina, G., Buhrmann, T., Neves, L., & Barbieri, F. (2022). *TweetNLP: Cutting-Edge Natural Language Processing for Social Media* (No. arXiv:2206.14774). arXiv. https://doi.org/10.48550/arXiv.2206.14774

Cardiff NLP. (2022, November 30). *Cardiffnlp/twitter-xlm-roberta-base-sentiment-multilingual*. Hugging Face. https://huggingface.co/cardiffnlp/twitter-xlm-roberta-base-sentiment-multilingual

Cedden, G., Meyer, P., Özkara, B., & Stutterheim, C. von. (2024). The “code-switching issue”: Transition from (socio)linguistic to cognitive research. *Bilingualism: Language and Cognition*, 1–14. https://doi.org/10.1017/S1366728924000737

Chatterjere, A., Guptha, V., Chopra, P., & Das, A. (2020). Minority positive sampling for switching points—An anecdote for the code-mixing language modeling. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 6228–6236). European Language Resources Association. https://aclanthology.org/2020.lrec-1.764

Chomsky, N. (1968). Quine’s Empirical Assumptions. *Synthese*, *19*(1/2), 53–68.

Deshpande, A., Murahari, V., Rajpurohit, T., Kalyan, A., & Narasimhan, K. (2023). Toxicity in chatgpt: Analyzing persona-assigned language models. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 1236–1270). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.88

Deuchar, M. (2010). *BilingBank Spanish-English Bangor Miami Corpus* [Dataset]. TalkBank. https://doi.org/10.21415/T5J01D

Dewaele, J.-M. (2010). *Emotions in multiple languages*. Palgrave Macmillan.

Doğruöz, A. S., Sitaram, S., Bullock, B. E., & Toribio, A. J. (2021). A Survey of Code-switching: Linguistic and Social Perspectives for Language Technologies. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 1654–1666). Association for

Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.131

Dunn, O. J. (1961). Multiple Comparisons among Means. *Journal of the American Statistical Association*, *56*(293), 52–64. https://doi.org/10.1080/01621459.1961.10482090

Gambäck, B., & Das, A. (2016). Comparing the level of code-switching in corpora. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 1850–1855). European Language Resources Association (ELRA). https://aclanthology.org/L16-1292

García-Vega, M., Díaz-Galiano, M., García-Cumbreras, M., Plaza-Del-Arco, F. M., Montejo-Ráez, A., Zafra, S. M., Martínez-Cámara, E., Aguilar, C., Antonio, M., Cabezudo, S., Chiruzzo, L., & Moctezuma, D. (2020). *Overview of TASS 2020: Introducing Emotion Detection*.

Gardner-Chloros, P. (2009). *Code-switching*. Cambridge University Press. https://doi.org/10.1017/CBO9780511609787

Gibadullin, I., Valeev, A., Khusainova, A., & Khan, A. (2019). *A Survey of Methods to Leverage Monolingual Data in Low-resource Neural Machine Translation* (No. arXiv:1910.00373). arXiv. https://doi.org/10.48550/arXiv.1910.00373

Grieve, J., Bartl, S., Fuoli, M., Grafmiller, J., Huang, W., Jawerbaum, A., Murakami, A., Perlman, M., Roemling, D., & Winter, B. (2024). *The Sociolinguistic Foundations of Language Modeling* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2407.09241

Hedderich, M. A., Lange, L., Adel, H., Strötgen, J., & Klakow, D. (2021). A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios. In K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R.

Cotterell, T. Chakraborty, & Y. Zhou (Eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 2545–2568). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.201

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). *spaCy: Industrial-strength Natural Language Processing in Python*. https://doi.org/10.5281/zenodo.1212303

Hovy, D. (2015). Demographic Factors Improve Classification Performance. In C. Zong & M. Strube (Eds.), *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 752–762). Association for Computational Linguistics. https://doi.org/10.3115/v1/P15-1073

Huang, Y., Yuan, Z., Zhou, Y., Guo, K., Wang, X., Zhuang, H., Sun, W., Sun, L., Wang, J., Ye, Y., & Zhang, X. (2024). *Social Science Meets LLMs: How Reliable Are Large Language Models in Social Simulations?* (No. arXiv:2410.23426). arXiv. https://doi.org/10.48550/arXiv.2410.23426

Hutto, C., & Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1), Article 1. https://doi.org/10.1609/icwsm.v8i1.14550

IMDb Benchmark (Sentiment Analysis). (2025). Papers with Code. https://paperswithcode.com/sota/sentiment-analysis-on-imdb

Jiawa, Z., Wei, L., Sili, W., & Heng, Y. (2021). Review of Methods and Applications of Text Sentiment Analysis. *Data Analysis and Knowledge Discovery*, *5*(6), Article 6. https://doi.org/10.11925/infotech.2096-3467.2021.0040

Jin, Y., Cheng, K., Wang, X., & Cai, L. (2023). A review of text sentiment analysis methods and applications. *Frontiers in Business, Economics and Management*, *10*(1), 58–64. https://doi.org/10.54097/fbem.v10i1.10171

Jose, N., Chakravarthi, B. R., Suryawanshi, S., Sherly, E., & McCrae, J. P. (2020). A Survey of Current Datasets for Code-Switching Research. *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, 136–141. https://doi.org/10.1109/ICACCS48705.2020.9074205

Kruskal, W. H., & and Wallis, W. A. (1952). Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association*, *47*(260), 583–621. https://doi.org/10.1080/01621459.1952.10483441

Labov, W. (1972). *Sociolinguistic Patterns*. University of Pennsylvania press.

Laurer, M., Atteveldt, W. van, Casas, A., & Welbers, K. (2024). Less Annotating, More Classifying: Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT-NLI. *Political Analysis*, *32*(1), 84–100. https://doi.org/10.1017/pan.2023.20

Leeman, J., & Fuller, J. M. (2021). *Hablar español en Estados Unidos: La sociopolítica del lenguaje*. Multilingual Matters.

Li, J. (2024). Area under the ROC Curve has the most consistent evaluation for binary classification. *PLOS ONE*, *19*(12), e0316019. https://doi.org/10.1371/journal.pone.0316019

Li, J., Yang, Y., Bai, Y., Zhou, X., Li, Y., Sun, H., Liu, Y., Si, X., Ye, Y., Wu, Y., 林一冠林一冠, Xu, B., Bowen, R., Feng, C., Gao, Y., & Huang, H. (2024). Fundamental Capabilities of Large Language Models and their Applications in Domain Scenarios: A Survey. In L.-W.

Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 11116–11141). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.599

Lik Xun Yuan. (2023). *Distilbert-base-multilingual-cased-sentiments-student*. https://doi.org/10.57967/HF/1422

LinCE Leaderboard. (2025). LinCE Benchmark. https://ritual.uh.edu/lince/leaderboard

Lipski, J. M. (2008). Spanish, English, or ... Spanglish? In *Varieties of Spanish in the United States* (pp. 38–74). Georgetown University Press. https://catalog.lib.unc.edu/catalog/UNCb5772376

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning Word Vectors for Sentiment Analysis. In D. Lin, Y. Matsumoto, & R. Mihalcea (Eds.), *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 142–150). Association for Computational Linguistics. https://aclanthology.org/P11-1015

Mabokela, K. R., Celik, T., & Raborife, M. (2023). Multilingual Sentiment Analysis for Under-Resourced Languages: A Systematic Review of the Landscape. *IEEE Access*, *11*, 15996–16020. https://doi.org/10.1109/ACCESS.2022.3224136

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk: Transcription format and programs, Vol. 1, 3rd ed* (pp. xi, 366). Lawrence Erlbaum Associates Publishers.

MacWhinney, B. (2019). *CHAT Manual*. TalkBank. https://doi.org/10.21415/3MHN-0Z89

MacWhinney, B. (2020). TalkBank for SLA. In *The Routledge Handbook of Second Language Acquisition and Corpora*. Routledge.

Mathur, S., & Shrivastava, G. (2024). Performance Analysis using Machine Learning for Code Mixed Languages in Sentiment Analysis. *International Journal on Advances in Engineering, Technology and Science (IJAETS)*, *5*(1), 99–104. https://doi.org/10.5281/zenodo.10719323

Myers-Scotton, C. (1993). *Duelling Languages: Grammatical Structure in Codeswitching*. Oxford University PressOxford. https://doi.org/10.1093/oso/9780198240594.001.0001

Nazir, M. K., Faisal, C. N., Habib, M. A., & Ahmad, H. (2025). Leveraging Multilingual Transformer for Multiclass Sentiment Analysis in Code-Mixed Data of Low-Resource Languages. *IEEE Access*, *13*, 7538–7554. IEEE Access. https://doi.org/10.1109/ACCESS.2025.3527710

Ochs, E. (1979). Planned and Unplanned Discourse. In *Discourse and Syntax* (pp. 51–80). Brill. https://doi.org/10.1163/9789004368897_004

OpenAI. (2024a, July 18). *GPT-4o mini: Advancing cost-efficient intelligence*. OpenAI. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

OpenAI. (2024b, August 8). *GPT-4o System Card*. OpenAI. https://openai.com/index/gpt-4o-system-card/

Pavlenko, A. (2008). Emotion and emotion-laden words in the bilingual lexicon. *Bilingualism: Language and Cognition*, *11*(2), 147–164. https://doi.org/10.1017/S1366728908003283

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830.

Pérez, J. M., Furman, D. A., Alonso Alemany, L., & Luque, F. M. (2022). RoBERTuito: A

pre-trained language model for social media text in Spanish. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 7235–7243). European Language Resources Association. https://aclanthology.org/2022.lrec-1.785

Pérez, J. M., Rajngewerc, M., Giudici, J. C., Furman, D. A., Luque, F., Alemany, L. A., & Martínez, M. V. (2024). *pysentimiento: A Python Toolkit for Opinion Mining and Social NLP tasks* (No. arXiv:2106.09462). arXiv. https://doi.org/10.48550/arXiv.2106.09462

Ranathunga, S., Lee, E.-S. A., Skenduli, M. P., Shekhar, R., Alam, M., & Kaur, R. (2021). *Neural Machine Translation for Low-Resource Languages: A Survey* (No. arXiv:2106.15115). arXiv. https://doi.org/10.48550/arXiv.2106.15115

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161–1178. https://doi.org/10.1037/h0077714

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). *DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter* (No. arXiv:1910.01108). arXiv. https://doi.org/10.48550/arXiv.1910.01108

Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., Liu, A., Si, C., Li, Y., Gupta, A., Han, H., Schulhoff, S., Dulepet, P. S., Vidyadhara, S., Ki, D., Agrawal, S., Pham, C., Kroiz, G., Li, F., Tao, H., Srivastava, A., … Resnik, P. (2025). *The Prompt Report: A Systematic Survey of Prompt Engineering Techniques* (No. arXiv:2406.06608). arXiv. https://doi.org/10.48550/arXiv.2406.06608

Sitaram, S., Chandu, K. R., Rallabandi, S. K., & Black, A. W. (2020). *A Survey of Code-switched Speech and Language Processing* (No. arXiv:1904.00784). arXiv.

https://doi.org/10.48550/arXiv.1904.00784

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013). Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In D. Yarowsky, T. Baldwin, A. Korhonen, K. Livescu, & S. Bethard (Eds.), *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (pp. 1631–1642). Association for Computational Linguistics. https://aclanthology.org/D13-1170/

Srinivasan, R., & Subalalitha, C. N. (2023). Sentimental analysis from imbalanced code-mixed data using machine learning approaches. *Distributed and Parallel Databases*, *41*(1), 37–52. https://doi.org/10.1007/s10619-021-07331-4

Stahl, P. M. (2025). *Pemistahl/lingua-py* [Python]. https://github.com/pemistahl/lingua-py (Original work published 2021)

Sutar, R., Kamalakar, D., & Desai, K. (2023). *A Study on Various Sentiment Analysis for Mixed Transliterated Indigenous Language using Machine Learning Algorithms*.

Tagliamonte, S. A. (2015). *Making waves: The story of variationist sociolinguistics* (1st ed.). Wiley. https://doi.org/10.1002/9781118455494

Thelwall, M., Buckley, K., & Paltoglou, G. (2012). Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, *63*(1), 163–173. https://doi.org/10.1002/asi.21662

Tseng, Y.-M., Huang, Y.-C., Hsiao, T.-Y., Chen, W.-L., Huang, C.-W., Meng, Y., & Chen, Y.-N. (2024). *Two Tales of Persona in LLMs: A Survey of Role-Playing and Personalization* (No. arXiv:2406.01171). arXiv. https://doi.org/10.48550/arXiv.2406.01171

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., &

Polosukhin, I. (2023). *Attention Is All You Need* (No. arXiv:1706.03762). arXiv. https://doi.org/10.48550/arXiv.1706.03762

Vilares, D., Alonso, M. A., & Gómez-Rodríguez, C. (2015). Sentiment Analysis on Monolingual, Multilingual and Code-Switching Twitter Corpora. In A. Balahur, E. van der Goot, P. Vossen, & A. Montoyo (Eds.), *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (pp. 2–8). Association for Computational Linguistics. https://doi.org/10.18653/v1/W15-2902

Vilares, D., Alonso, M. A., & Gómez-Rodríguez, C. (2016). EN-ES-CS: An English-Spanish Code-Switching Twitter Corpus for Multilingual Sentiment Analysis. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 4149–4153). European Language Resources Association (ELRA). https://aclanthology.org/L16-1655

Volkova, S., Wilson, T., & Yarowsky, D. (2013). Exploring Demographic Language Variations to Improve Multilingual Sentiment Analysis in Social Media. In D. Yarowsky, T. Baldwin, A. Korhonen, K. Livescu, & S. Bethard (Eds.), *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (pp. 1815–1827). Association for Computational Linguistics. https://aclanthology.org/D13-1187/

Wang, F., Zhang, Z., Zhang, X., Wu, Z., Mo, T., Lu, Q., Wang, W., Li, R., Xu, J., Tang, X., He, Q., Ma, Y., Huang, M., & Wang, S. (2024). *A Comprehensive Survey of Small Language Models in the Era of Large Language Models: Techniques, Enhancements, Applications, Collaboration with LLMs, and Trustworthiness* (No. arXiv:2411.03350). arXiv. https://doi.org/10.48550/arXiv.2411.03350

Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, *55*(7), 5731–5780. https://doi.org/10.1007/s10462-022-10144-1

Waugh, L. R., Monville-Burston, M., & Joseph, J. E. (Eds.). (2023). *The Cambridge History of Linguistics*. Cambridge University Press. https://doi.org/10.1017/9780511842788

Winata, G., Aji, A. F., Yong, Z. X., & Solorio, T. (2023). The Decades Progress on Code-Switching Research in NLP: A Systematic Survey on Trends and Challenges. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Findings of the Association for Computational Linguistics: ACL 2023* (pp. 2936–2978). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.185

Yang, K., Li, H., Wen, H., Peng, T.-Q., Tang, J., & Liu, H. (2024). *Are Large Language Models (LLMs) Good Social Predictors?* (No. arXiv:2402.12620). arXiv. https://doi.org/10.48550/arXiv.2402.12620

Zeng, L. (2024). *Leveraging Large Language Models for Code-Mixed Data Augmentation in Sentiment Analysis* (No. arXiv:2411.00691). arXiv. https://doi.org/10.48550/arXiv.2411.00691

Zhang, W., Deng, Y., Liu, B., Pan, S. J., & Bing, L. (2023). *Sentiment Analysis in the Era of Large Language Models: A Reality Check* (No. arXiv:2305.15005). arXiv. https://doi.org/10.48550/arXiv.2305.15005

Appendix

Language Models

	SLM / LLM	*Model details*
LM1	SLM	pysentimiento/robertuito-sentiment-analysis (García-Vega et al., 2020; Pérez et al., 2022, 2024) 109M parameters
LM2	SLM	lxyuan/distilbert-base-multilingual-cased-sentiments-student (Laurer et al., 2024; Lik Xun Yuan, 2023; Sanh et al., 2020) 135M parameters
LM3	SLM	cardiffnlp/twitter-xlm-roberta-base-sentiment-multilingual (Camacho-Collados et al., 2022; Cardiff NLP, 2022) 279M parameters
LM4	LLM	GPT-4o mini (OpenAI, 2024a) Estimated 8B parameters
LM5	LLM	GPT-4o (OpenAI, 2024b) Estimated 200B parameters

ULSM Equation

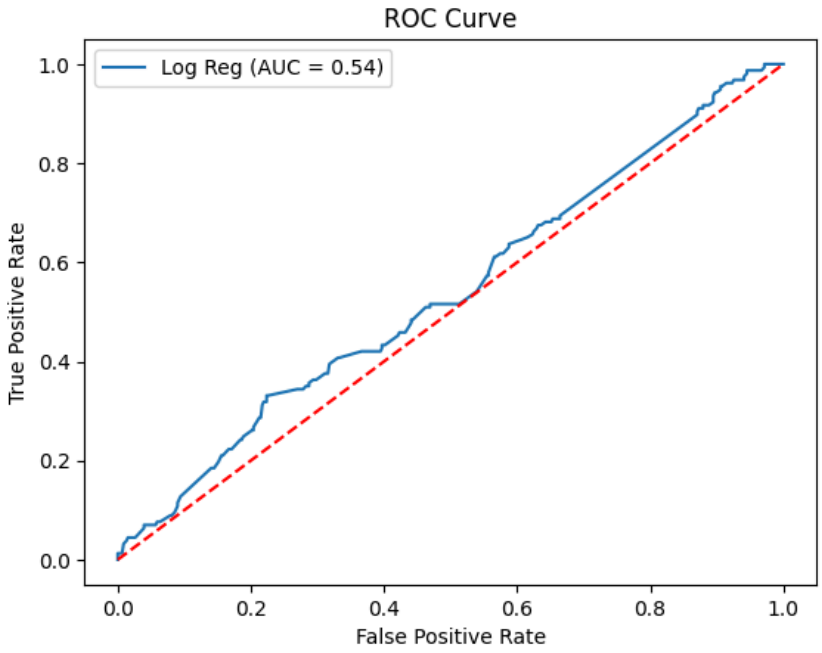
$$\text{ULSM} = 100 \cdot \frac{w_m \left(N - \max_{L_i \in \mathbb{L}} \{t_{L_i}\} \right) + w_p P}{N}$$

Where:

- N is the number of tokens in the utterance that correspond to some language,
- t is the number of tokens in the utterance that correspond to the matrix language,
- P is the number of code alternation points, and
- w_m, w_p are weights for the matrix and the code alternation points such that $w_m + w_p = 1$

From Gambäck and Das (2016). For our experiments, I set $w_m = w_p = 0.5$, and treated the utterance-level majority language as the matrix language. Note that this differs from traditional matrix language-based syntactic accounts of code-switching; see Myers-Scotton (1993).

ULSM Logistic Regression ROC Curve



The ROC (receiver operating characteristic) curve visualizes the predictive capacity of a binary model. The independent variable is false positive rate, and the independent variable is true positive rate. The area under this ROC curve, 0.54, is an average-case indication of the system's classification ability.

LLM Prompts

CS-naive prompts

Zero-shot prompt

You are a helpful and unbiased language model. *Classify the sentiment of the following sentence as "Positive", "Negative", or "Neutral."* *Do not provide any explanation or additional text in your response; output only the sentiment label.*

Few-shot prompt

You are a helpful and unbiased language model.

Here are some examples of sentiment classification:

Example 1:
Sentence: I love this product!
Sentiment: Positive

Example 2:
Sentence: I hate this service.
Sentiment: Negative

Example 3:
Sentence: It's okay, I guess.
Sentiment: Neutral

Now, classify the sentiment of the following sentence in the same style using only one of these labels: Positive, Negative, or Neutral.
Do not provide any explanation or additional text—only the label.

Chain-of-thought prompt

You are a helpful and unbiased language model.

Please analyze the sentiment of the following sentence step by step.
Explain your reasoning (chain-of-thought), and then clearly provide your final answer.

Your chain-of-thought might look like this:
1. Identify any words or phrases that carry emotional weight.
2. Consider the context and overall tone.
3. Decide which sentiment category (Positive, Negative, or Neutral) best fits.

Finally, write:
Chain-of-thought: <your step-by-step reasoning>
Final classification: <Positive/Negative/Neutral>

CS-aware prompts

Zero-shot prompt

You are a helpful and unbiased language model.
You are capable of understanding and classifying mixed-language sentences, where different

languages may appear within the same sentence.
Classify the sentiment of the following sentence as "Positive", "Negative", or "Neutral."

Please ensure that you handle any code-switching appropriately by considering the context and emotions in all languages used.
Do not provide any explanation or additional text in your response; output only the sentiment label.

Few-shot prompt

You are a helpful and unbiased language model.
Here are some examples of sentiment classification for sentences with mixed languages (code-switching):

Example 1:
Sentence: Me encanta este producto! Es muy útil.
Sentiment: Positive

Example 2:
Sentence: No me gusta este servicio, it's terrible.
Sentiment: Negative

Example 3:
Sentence: Está bien, I guess.
Sentiment: Neutral

Now, classify the sentiment of the following sentence in the same style, accounting for any code-switching between languages, using only one of these labels: Positive, Negative, or Neutral.
Do not provide any explanation or additional text—only the label.

Chain-of-thought prompt

You are a helpful and unbiased language model.

Please analyze the sentiment of the following sentence step by step, keeping in mind that the sentence may contain code-switching between languages.
Explain your reasoning (chain-of-thought), and then clearly provide your final answer.

Your chain-of-thought might look like this:
1. Identify any words or phrases in the sentence that carry emotional weight, regardless of language.
2. Consider the context and overall tone, taking into account the emotions expressed in all languages used.

3. Decide which sentiment category (Positive, Negative, or Neutral) best fits the sentence, based on the full mixed-language context.

Finally, write:
Chain-of-thought: <your step-by-step reasoning>
Final classification: <Positive/Negative/Neutral>