The SlangTrack Dataset: Supporting the Detection of Words Used in Slang Senses

Afnan Aloraini[1,2], **Goran Nenadic**[1], **Viktor Schlegel**[1] **and Riza Batista-Navarro**[1]
[1]Department of Computer Science, The University of Manchester, UK
[2]Department of Computer Science, Qassim University, Saudi Arabia
{afnan.aloraini, g.nenadic, viktor.schlegel, riza.batista}@manchester.ac.uk
A.ALOURANI@qu.edu.sa

Abstract

Slang is widespread in informal communication, yet its fluidity poses challenges for natural language processing (NLP), especially when words alternate between slang and non-slang senses. While prior work has examined slang through dictionaries, sentiment analysis, and lexicon building, little attention has been given to detecting slang usage in context. We address this gap by re-framing slang detection as distinguishing slang from non-slang senses of the same lexical item. To support this task, we introduce *SlangTrack* (ST), a diachronically structured dataset of dual-meaning words annotated at the sentence level with high inter-annotator agreement. We benchmark (1) deep learning models with static and contextual embeddings, (2) transformer-based models, and (3) large language models evaluated in zero-shot, few-shot, and fine-tuned settings. Fine-tuned transformers, especially BERT-large enriched with sentiment and emotion features, achieve the strongest performance, reaching an F1-score of 72% for slang and 92% for non-slang usage. Our findings highlight both the difficulty of contextual slang detection and the value of affective cues for improving model robustness.

1 Introduction

Disclaimer: This work includes offensive slang examples, which do not reflect the researchers' views.

Slang is an informal linguistic phenomenon comprising words and phrases used within specific groups (Dumas and Lighter, 1978; Adams, 2012; Green, 2015; Sun et al., 2021). Although linguists and computer scientists have developed many methods for semantic analysis, slang remains a persistent challenge for natural language processing (NLP) systems (Eisenstein, 2013; Blodgett et al., 2016; Wuraola et al., 2024). It often encodes emotions, attitudes, and social affiliation elements that are crucial for interpreting meaning in context (Sandow et al., 2024; Haber and Poesio, 2024).

Identifying slang is important for tracking semantic change and improving NLP tasks (Adams, 2012; Sun et al., 2022; Keidar et al., 2022; Sun et al., 2024). Many terms evolve rapidly, for example, "cool" shifted from a literal descriptor to an evaluative slang term (Coleman, 2012; Dhuliawala et al., 2016). Accurate disambiguation of such meanings is essential for applications like conversational agents, machine translation, and sentiment analysis, where figurative senses are often misinterpreted.

Detecting slang is challenging because many terms have multiple meanings depending on context, often resulting in double entendres (Kiddon and Brun, 2011). For example, "He's a player on Sundays" can refer to sports or manipulative romantic behaviour, highlighting the need for contextual disambiguation. Recent work in NLP has begun to recognise the importance of informal and slang language as meaningful phenomena for modern language technologies. For example, benchmarks and evaluation frameworks have been proposed to measure models' capabilities in processing slang and other informal language forms using neural methods (Sun et al., 2024; Pei et al., 2019). Studies of large language models also show that slang remains challenging for foundational models and serves as a useful probe of their linguistic knowledge (Sun et al., 2024).

Despite these emerging efforts, slang detection has received relatively limited attention in the broader NLP community. Existing resources tend to focus either on dictionary–text mismatches (Pei et al., 2019) or on scripted conversational data such as movie subtitles (Sun et al., 2024), and do not address words that appear in both slang and literal senses. We refer to these as *dual-meaning words*.

In this work, slang detection is framed as a context-sensitive and pragmatic distinction between slang and non-slang usage of the same lexical item, rather than a simple lexical lookup. Be-

The Proceedings for the 6th International Workshop on Computational Approaches to Language Change (LChange'26), pages 1–19

cause our analysis draws on both historical and contemporary corpora, it aligns with computational research on semantic change (Hamilton et al., 2016; Schlechtweg et al., 2020; Kutuzov et al., 2018; Periti et al., 2024). Many slang senses emerge as pragmatic extensions of earlier meanings, making dual-meaning words a useful lens for studying sense competition and semantic–pragmatic drift over time (de Sá et al., 2024; Tahmasebi et al., 2021). This framing is inspired by word sense disambiguation (WSD) research, without committing to full sense-level modelling, and allows register-level (slang vs. non-slang) annotations to be aligned across corpora in a controlled manner.

This paper introduces a binary classification system designed to determine whether a target word occurrence is used in a slang or non-slang (standard) sense within its given textual instance. Our primary research question is: *Can algorithms reliably distinguish slang usage from non-slang usage at the instance level?* For example, the word "*salty*" can be used literally ("the soup is salty") or figuratively to express resentment ("she was salty after losing"). We frame slang detection as an instance-level (target-word–centered) classification task, where each instance is annotated and evaluated individually. Each instance corresponds to the same textual input provided to both annotators and models: a single tweet (which may contain multiple sentences) in Twitter, or a paragraph-length context in CCOHA. This framing provides a controlled and comparable unit of analysis across corpora, while allowing naturally occurring pragmatic and contextual cues to be captured within the textual span itself. Our contributions include the following:

- SlangTrack, a new corpus annotated with slang and non-slang labels at the instance level, supporting binary classification of target-word usage.
- Benchmarking a range of slang detection models, including: (1) basic neural classification models, (2) fine-tuned transformer-based language models (LMs), and (3) large language models (LLMs) evaluated under fine-tuned, zero-shot, and few-shot settings.
- An analysis of the role of sentiment and emotion cues in distinguishing slang from non-slang usage.
- An error analysis of misclassified instances produced by the best-performing model.

2 Related work

2.1 Construction of Slang Dictionaries and Sentiment Analysis

The development of resources for processing informal language, particularly slang, has been central in computational linguistics. Early efforts focused on constructing structured representations of slang, such as SlangNet (Dhuliawala et al., 2016), which integrates Urban Dictionary [1] entries into a WordNet-style framework (Miller, 1995). SlangSD (Wu et al., 2018) extends this line of work by providing a continually updated sentiment lexicon for slang terms derived from large-scale web and social media data. SLANGZY (Gupta et al., 2019) further applies machine learning to dynamically interpret slang and support downstream applications such as chatbots and social media analytics. Although these resources expand lexical and sentiment coverage, they do not annotate contextual usage or distinguish between non-slang and slang senses of the same word, which is central to our task.

2.2 Slang Word Creation and Interpretation

Research on slang detection and interpretation has evolved alongside the growth of digital communication. Early studies constructed slang lexicons (Pal and Saha, 2015), while later work employed neural architectures such as BiLSTMs, CRFs, and MLPs for classification (Pei et al., 2019). Beyond detection, sequence-to-sequence models have been used to interpret non-standard English (Ni and Wang, 2017), and BiLSTM- and GRU-based systems have shown strong performance in identifying domain-specific slang (Lynn et al., 2019). More recently, the OpenSub-Slang dataset (Sun et al., 2024) introduced demographic, contextual, and historical metadata for 7,488 slang-related sentences, enabling more nuanced modelling. Additional work combining semantic and contextual cues (Sun et al., 2022) has further advanced slang interpretation.

Recent developments in WSD are also relevant to our task. Models such as GlossBERT (Huang et al., 2019), EWISE (Kumar et al., 2019), and EWISER (Bevilacqua and Navigli, 2020) demonstrate the effectiveness of supervised context–gloss methods for resolving semantic ambiguity. However, these approaches do not target slang senses specifically or model the contrast between slang and literal usage, which remains understudied.

[1] https://www.urbandictionary.com

2.3 Research Gaps

Although prior work has advanced slang detection, it has generally treated slang as a broad category rather than focusing on dual-meaning words. For example, Pei et al. (2019) contrasts dictionary-derived slang with formal news text, and Sun et al. (2024) captures conversational slang using movie subtitles. Other studies draw on Urban Dictionary or focus on specific domains such as misogynistic slang (Lynn et al., 2019). More recently, the OpenSubtitles-Slang dataset introduced cross-lingual mappings for slang terms, providing paraphrases or equivalents in multiple languages (Sun et al., 2024). However, in these resources, slang and non-slang items originate from different corpora, so the same lexical item is not annotated with both slang and non-slang senses. As a result, the task becomes identifying slang words as a category rather than determining whether a familiar word is used in a slang or a non-slang sense in context.

Our study addresses this gap by re-framing slang detection as an instance-level contextual disambiguation problem. Instead of asking whether a text contains slang expressions, we ask whether a specific occurrence of a polysemous word is used with a slang (socially marked/figurative) or non-slang (literal/standard) meaning in context. This perspective is inspired by word sense disambiguation, but focuses on register-level distinctions, extending prior slang research to an underexplored setting in which the same lexical item alternates between slang and non-slang usage.

3 Problem Formulation

We frame slang detection as an instance-level classification task inspired by word sense disambiguation, applied to polysemous dual-meaning words, each with at least one slang and one non-slang sense. Let w denote such a target word and s a sentence containing w. Each instance is annotated with a binary label indicating whether its sense in context is `slang` or `non-slang`, and the task is to assign the correct label $y \in \{\texttt{slang}, \texttt{non-slang}\}$ to each occurrence of w.

The word *salty* illustrates the range of senses we consider: a literal taste meaning ("the soup is salty"), slang for resentment ("he was salty after losing"), slang for old or worn ("a salty jacket"), slang for toughness ("a salty veteran"), slang for vulgarity ("a salty bar"), and non-slang proper-noun uses ("Salty is the name of their dog"). These senses demonstrate how a single lexical item spans multiple non-slang and slang sub-senses; in SlangTrack, this variability is intentionally collapsed into a binary decision, making contextual disambiguation of socially marked (slang) versus non-slang usage central to the task. This formulation prioritises register-level discrimination over fine-grained sense distinctions.

4 Dataset

Existing resources for slang research face limitations for binary slang classification and polysemy-sensitive analysis. Dictionary-based resources such as Urban Dictionary (Ni and Wang, 2017), the Online Slang Dictionary (OSD), [2] and Green's Dictionary of Slang (GDoS) (Adams, 2012) offer broad lexical coverage but lack systematically annotated non-slang counterparts and do not capture contextual usage. Reddit glossaries (Lucy and Bamman, 2021) provide community-specific slang terms, but remain at the glossary level rather than sentence level.

Two datasets support binary slang classification: Pei et al. (2019) and OpenSub-Slang (Sun et al., 2024). Pei et al. combine dictionary-derived slang with negative samples drawn heuristically from news text, raising the risk of domain artefacts. OpenSub-Slang offers richer contextual information, including paraphrases and demographic metadata, but its reliance on scripted dialogue introduces domain biases and limits generalisation to real-world slang. Neither resource addresses dual-meaning words that alternate between slang and non-slang meanings. A comparison of slang-related datasets is shown in Table 1.

To address these gaps, we introduce the SlangTrack (ST) dataset.[3] The ST dataset is built around dual-meaning target words, each possessing at least one slang and one non-slang sense. Unlike dictionary-based or scripted sources, ST draws on naturally occurring language in both the Cleaned Corpus of Historical American English (CCOHA) (Alatrash et al., 2020) and contemporary Twitter.[4] This combination provides coverage of modern slang usage while enabling contextual

[2] https://www.onlineslangdictionary.com

[3] Publicly available at: https://github.com/SlangTrack/SlangTrack-ST-/blob/main/README.md

[4] https://twitter.com. Our dataset includes social media text, which may contain offensive material. All excerpts are de-identified in accordance with Twitter's Academic Research TOS.

Dataset	Source / Domain	Temp. Cov.	Slang Cov.	Non-slang Cov.	Polysemy (dual-meaning)	Annotation	Fully Annot.	Bench.	Publicly Available
Urban Dictionary (Ni and Wang, 2017)	Crowdsourced slang dictionary (defs. + examples)	Contemp.	✓	✗	✗	Word / entry-level	✗	✗	✓
Online Slang Dictionary (OSD) (Sun et al., 2022)	Curated slang dictionary (defs. + examples)	Contemp.	✓	✗	✗	Word / entry-level	✗	✗	✗
Green's Dictionary of Slang (GDoS) (Adams, 2012)	Historical lexicographic dictionary (defs. + citations)	Hist.	✓	✗	✗	Word / entry-level	✗	✗	✗
Reddit Glossaries (Lucy and Bamman, 2021)	Community slang glossaries (subreddit lists)	Contemp.	✓	✗	✗	Word / glossary-level	✗	✗	✓
OpenSubtitles-Slang (Sun et al., 2024)	Scripted dialogue (movie subtitles)	Contemp.	✓	✓	✗	Sentence-level (+ slang tokens, subset annotated)	✗	✓	✓
SlangTrack (ST)	Naturally occurring text in two diverse corpora	Both	✓	✓	✓	Instance-level (target-word–centered; slang vs. non-slang)	✓	✓	✓

Table 1: Comparison of slang-related datasets. "*Temp. Cov.*" = temporal coverage ("*Hist.*" pre-2000; "*Contemp.*" post-2000; "*Both*" spans both). "*Slang / Non-slang Cov.*" = slang vs. non-slang usage. "*Polysemy*" = dual meanings. "*Annotation*" = granularity. "*Fully Annot.*" = fully gold-labelled. "*Bench.*" = benchmark.

disambiguation across time and registers.

4.1 Data Collection (Target Words and Examples)

We selected target words that appeared in both the SlangSD wordlist [5] and CCOHA, prioritising items with multiple attested senses. Each chosen word has at least one slang and one non-slang sense. To compile sense inventories, we consulted Green's Dictionary of Slang, Urban Dictionary, and the Online Slang Dictionary, and cross-referenced meanings with the Oxford English Dictionary (OED) [6] to confirm non-slang usage. Full sense inventories are provided in Appendix 12.

We ensured that each word appeared in both CCOHA and Twitter, allowing the dataset to reflect historical and contemporary senses. Applying these criteria yielded ten target words. Although ST contains only ten target words, this limited lexical scope is intentional and methodologically motivated. Each selected word exhibits between two and eight attested senses, requiring careful cross-referencing across slang dictionaries, OED entries, and corpus attestations. Because contextual disambiguation for dual-meaning words is annotation-intensive, especially for long and paraphrastic CCOHA sentences, focusing on a smaller set ensured high annotation quality and interpretability of sense contrasts (see Appendix 16). This design follows established practice in semantic change research, where controlled, high-quality pilot datasets with narrow lexical coverage (e.g., SemEval 2020 Task 1) serve as foundational benchmarks before broader scaling. Accordingly, ST itself should be interpreted as a controlled pilot dataset, prioritising sense precision and annotation quality over breadth. This initial release establishes a reliable foundation that we will expand in future work with additional words, platforms, and more recent data. Each instance is labelled with its source (CCOHA or Twitter) and timestamp (CCOHA publication year or tweet date). An instance corresponds to the naturally occurring textual unit in the source corpus: a single tweet in Twitter or a paragraph-length context in CCOHA. While instances are evaluated independently, their length varies across sources, allowing some pragmatic and contextual cues to be captured without explicitly modelling broader discourse history. Although the task centres on slang detection rather than temporal modelling, this metadata enables users to distinguish older formal contexts from modern informal ones. While final labels are binary (slang vs. non-slang), annotators relied on the predefined sense inventory to determine whether each instance corresponded to an attested slang or non-slang sense. This ensures that the slang label reflects a genuine register distinction grounded in attested meanings, rather than surface stylistic informality.[7]

[5] https://rdocumentation.org/packages/lexicon

[6] https://www.oed.com

[7] We also release a fine-grained sense-labelled version, SlangTrack-WSD (ST-WSD), derived from the same annotation process. While ST-WSD preserves fine-grained sense distinctions, the present study focuses on binary register-level disambiguation; a direct comparison between binary and

Taken together, these design decisions give ST synchronic labels over a diachronic dataset, where slang and non-slang senses show distinct patterns across CCOHA and Twitter. Even without explicit temporal modelling, this structure enables analysis of how socially marked senses vary across time and registers.

CCOHA contributed historical usage spanning 1980–2010, including non-slang and occasional slang contexts. For contemporary usage, we collected up to 1,000 tweets per target word from 2010–2020 through the Twitter API. Combining both corpora allows ST to capture emerging slang from social media alongside formal usage from CCOHA, balancing sources and enabling direct comparison of dual-meaning words across registers. We deliberately restrict our sources to CCOHA and Twitter to create a clear contrast between historical formal writing and contemporary informal usage.

4.2 Annotation Guidelines and Details

Annotators were provided with the target words, example sentences, and a predefined sense inventory. Using this information, they labelled each instance as *slang* or *non-slang* based on whether the usage matched an attested slang sense or a non-slang sense.[8] All instances were mapped to a binary slang versus non-slang classification setting.

The annotation team consisted of three English-proficient annotators, one with a linguistics background who served as the primary annotator. Prior to full-scale annotation, an initial pilot phase was conducted to refine the annotation guidelines. Following this, two annotators independently labelled all instances in the dataset (12,712 instances) using the finalised guidelines.

Inter-annotator agreement was computed on the independent (pre-adjudication) labels using Cohen's Kappa and reached 0.887, indicating high overall reliability. Disagreements were subsequently resolved by the primary annotator through adjudication. The resulting class distribution is skewed, with approximately 80% non-slang and 20% slang instances. This imbalance reflects naturally occurring usage patterns rather than annotation or sampling decisions. To account for class imbalance, we report macro-averaged evaluation metrics and conduct per-word analyses rather than relying on accuracy alone.

4.3 Data Statistics

The dataset contains 12,712 labelled instances: 10,105 non-slang and 2,607 slang. It includes 48,508 unique word types and 310,170 tokens. The average post length is 34.6 words and contains 3.74 sentences. Using stratified sampling, the dataset was divided into 70% training, 15% validation, and 15% test splits.

Keyword	Non-slang	Slang	Total
BMW	1,082	15	1,097
Brownie	706	258	964
Chronic	1,259	426	1,685
Climber	505	137	642
Cucumber	978	73	1,051
Eat	2,708	324	3,032
Germ	753	79	832
Mammy	904	166	1,070
Rodent	744	329	1,073
Salty	535	731	1,266
Total	10,105	2,607	12,712

Table 2: Distribution of slang and non-slang instances per target word.

Breaking this down by corpus, CCOHA contributes 3,660 non-slang and 756 slang examples, while Twitter contributes 6,514 non-slang and 1,782 slang examples. The resulting class distribution (approximately 80% non-slang and 20% slang) reflects naturally occurring usage patterns in the underlying corpora rather than annotation or sampling decisions.

To explore stylistic variation between slang and non-slang usage, we conducted a one-way ANOVA on automatically derived sentiment and emotion scores. Because these scores are produced by an external classifier, the analysis is exploratory and not used to draw causal conclusions or justify model performance. The results show statistically significant but small differences: slang instances tend to be slightly more negative and express anger and sadness, while non-slang instances are more neutral. Full descriptive statistics appear in Appendix 11.

sense-level labels is left to future work. ST-WSD is provided as a companion resource: https://github.com/SlangTrack/SlangTrack-Word-Sense-Disambiguation.

[8] Proper-noun usages (e.g., brand names, organization names, song titles) are grouped under the non-slang category. While these uses are not always semantic "senses" in a strict lexicographic sense, they are not socially marked or figurative in the way slang usages are. For the purposes of this task, which focuses on distinguishing slang from non-slang register rather than enumerating fine-grained sense inventories, treating proper-noun usages as non-slang provides a consistent and operational distinction.

4.4 Diachronic Variation in Slang and non-slang Senses

Because ST dataset includes both historical COHA data and contemporary Twitter data, it allows us to examine how the balance between slang and non-slang senses changes over time. In COHA, most target words appear predominantly in their literal senses (e.g., *Eat* <3% slang usage, *BMW* <6%, *Cucumber* <13%), while a few items such as *Rodent* and *Chronic* show moderate slang usage (around 25–30%). Twitter displays a different distribution. Some words become substantially more slang-dominant: *Salty*, for instance, shifts from being mostly literal in COHA to being used slang overwhelmingly in Twitter (>70%). Other items show the opposite pattern. *Mammy* appears frequently in slang senses in COHA, but these uses become rare in modern Twitter data. Several words, including *Eat*, *Climber*, and *Cucumber*, remain largely stable, with literal senses prevailing in both corpora. Although ST spans multiple historical periods, time is not modelled as an explicit predictive variable. Instead, it is treated as a latent property of the data: instances are timestamped, but classification relies solely on local textual context, allowing us to isolate instance-level slang disambiguation while still supporting descriptive temporal analysis distributed across periods.

A statistical comparison of slang proportions between COHA and Twitter (Appendix 13) confirms these trends. *Salty* and *Mammy* exhibit large, highly significant shifts ($p < .001$), while most other words show little or no significant change. Taken together, these results highlight that the dual-meaning words in ST follow diverse diachronic paths, illustrating patterns of sense innovation, decline, and stability over time.

5 Methodology

5.1 Pre-processing

We applied standard text pre-processing to reduce noise and ensure consistent input across models. This included removing URLs and usernames, excluding instances in which the target word appeared inside a URL, and eliminating duplicated content. Duplicate removal was performed at the exact string level, meaning that an instance was removed only if the entire textual content was repeated verbatim more than once in the dataset. Text normalisation steps such as lowercasing and punctuation removal were applied only in settings using static word embeddings with CNN and BiLSTM models, where such preprocessing is standard practice. For transformer-based and large language models, text was minimally processed, preserving original casing and punctuation in line with their pre-training regimes.

5.2 Evaluation

Models were evaluated using precision, recall, F1 (macro and weighted), and accuracy. Macro scores emphasise performance on the minority slang class, while weighted scores account for class imbalance. F1 captures the balance between precision and recall and is the primary metric for slang detection.

5.3 Classification Models

5.3.1 Basic classification models

We experimented with two neural architectures, Convolutional Neural Networks (CNNs) (LeCun and Bengio, 1995) and Bidirectional Long Short-Term Memory (BiLSTM) networks (Liu and Guo, 2019; Pei et al., 2019), for slang classification. In a first setting, we trained CNN and BiLSTM classifiers using pre-trained word embeddings only, allowing us to assess the contribution of different embedding types in isolation. Specifically, we experimented with three types of pre-trained embeddings as input features: FastText, GloVe (Pennington et al., 2014), and BERT-based contextual embeddings (Devlin et al., 2019). CNNs capture local lexical patterns, while BiLSTMs model longer-range sequential dependencies within each textual instance. In a second setting, we followed the architecture proposed by (Pei et al., 2019) and evaluated CNN–CRF and BiLSTM–CRF models using the full feature configuration [9] described in their work. This configuration combines word- and character-level representations with a set of linguistically motivated features, including part-of-speech (POS) tags, POS transition features, and pointwise mutual information (PMI) scores capturing atypical local word co-occurrence patterns associated with slang usage.

[9] Full features refers to the feature configuration proposed by Pei et al. (2019), which augments neural representations with linguistically motivated features such as part-of-speech (POS) information and pointwise mutual information (PMI) scores capturing atypical word co-occurrence patterns. We follow their implementation without modification.

5.3.2 Fine-tuning pre-trained transformer encoders

We fine-tuned four transformer models, BERT (Devlin et al., 2019), ALBERT (Lan, 2019), RoBERTa (Liu, 2019), and XLNet (Yang, 2019) using five-fold cross-validation over 30 epochs.[10] Although trained under the same protocol, the models differ architecturally: BERT uses masked bidirectional language modelling; ALBERT reduces parameters via factorisation and weight sharing; RoBERTa extends BERT with training; and XLNet uses permutation-based modelling for bidirectional context.

5.3.3 Fine-tuning large language models (LLMs)

We fine-tuned several Large Language Models (LLMs) for slang classification, including GPT-4o [11] (version GPT-4o-2024-08-06), a high-capacity model, and GPT-4o-mini [12] (version GPT-4o-mini-2024-07-18), a smaller and more cost-efficient variant. We also fine-tuned LLaMA-3.1-8B-Instruct and LLaMA-3.1-70B-Instruct [13], applying Low-Rank Adaptation (LoRA) to specialise these open-source models for slang detection. In addition to fine-tuning, we evaluated zero-shot (GPT-4o ZSp) and few-shot (GPT-4o FSp) prompting with structured task instructions.

5.4 Enhancing Slang Detection through Sentiment and Emotion Analysis

To capture affective cues in figurative slang, we incorporate automatically derived sentiment and emotion features into transformer-based and LLM models. Features were extracted from pre-trained emotion and sentiment classifiers [15] and concatenated with contextual embeddings from ALBERT-xxlarge-v2 and BERT-large-uncased prior to classification. Although contextual language models encode affect implicitly, explicitly modelling sentiment and emotion appears to provide a targeted inductive bias for socially marked and evaluative slang usages, which may otherwise be diluted in general-purpose representations. The impact of affective features is shown in Table 4. Full prompts and parameter settings for all models are provided in Tables 14 and 13.[16]

6 Results

Deep learning architectures provided competitive baselines for slang detection. BiLSTM models using BERT embeddings achieved the strongest performance, with an accuracy of 0.85, outperforming variants trained on GloVe or FastText. This confirms that contextualised embeddings better capture slang-related semantic variation than static representations. CNN models performed reasonably well but struggled with longer or more polysemous constructions, reflecting a limited ability to model sequential dependencies. Incorporating POS and PMI features (Pei et al., 2019) did not improve performance. BiLSTM-CRF and CNN-CRF models underperformed simpler BiLSTM variants, particularly on the minority slang class.

Transformer-based models produced stronger results overall. BERT-large-uncased achieved the highest accuracy of 0.87 and an F1 score of 0.69 on slang instances. These results highlight the advantage of large bidirectional models for contextual sense disambiguation. RoBERTa and XLNet also performed well but showed slightly lower slang recall, suggesting difficulty in distinguishing figurative slang uses from literal ones in ambiguous contexts. This aligns with prior findings that BERT-style models capture subtle semantic distinctions more reliably (Sun et al., 2024).

Large language models showed a similar pattern. Fine-tuned GPT-4o-mini reached an accuracy of 0.86 and slightly outperformed GPT-4o despite its smaller size. This indicates that fine-tuning enhances sensitivity to slang-related cues and allows compact models to rival BERT-large-uncased. Zero-shot and few-shot GPT-4o performed well on non-slang usage but showed reduced recall for slang, suggesting that prompting alone does not fully resolve dual-meaning ambiguity. LLaMA-3.1 models underperformed relative to GPT-4o and BERT-based systems. The 70B variant produced strong non-slang performance, with a precision of 0.87 and a recall of 0.64, but struggled with slang

[10] Each model was trained with three random seeds; performance variation was minimal (macro-F1 SD $\approx$ 0.01).

[11] https://platform.openai.com/docs/models/gpt-4o

[12] https://platform.openai.com/docs/models/gpt-4o-mini

[13] https://huggingface.co/meta-llama/Llama-3.1-8B

[14] Keras refers to embeddings created using https://www.tensorflow.org/api_docs/python/tf/keras/layers/Embedding, which are randomly initialised and trained during the model's learning process on task-specific data.

[15] Emotion embeddings from https://huggingface.co/bhadresh-savani/bert-base-go-emotion; sentiment embeddings from https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment.

[16] This applies to all classification models described in this section.

(a) Deep learning models

Models	Features	Non-slang			Slang			Avg. Macro Scores			Avg. Weighted Scores			
		Pr	Rec	F1	Pr	Rec	F1	Pr_M	Rec_M	F1_M	Pr_W	Rec_W	F1_W	Acc
BiLSTM	GloVe	0.88	0.92	0.90	0.69	0.56	0.62	0.78	0.74	0.76	0.83	0.84	0.83	0.84
	BERT	0.86	0.96	0.91	0.78	0.47	0.58	0.82	0.72	**0.75**	0.84	0.85	0.83	**0.85**
	FastText	0.88	0.91	0.90	0.66	0.58	0.62	0.77	0.75	0.76	0.83	0.84	0.83	0.84
BiLSTM-CRF (full features) (Pei et al., 2019)	Keras[14]	0.84	0.88	0.86	0.52	0.45	0.48	0.68	0.66	0.67	0.77	0.78	0.77	0.78
CNN	GloVe	0.87	0.87	0.87	0.61	0.61	0.61	0.74	0.74	0.74	0.81	0.81	0.81	0.81
	BERT	0.87	0.93	0.90	0.70	0.52	0.60	0.78	0.73	0.75	0.83	0.84	0.83	0.84
	FastText	0.88	0.91	0.90	0.66	0.58	0.62	0.77	0.75	**0.76**	0.83	0.84	0.83	**0.84**
CNN-CRF (full features) (Pei et al., 2019)	Keras	0.87	0.90	0.89	0.63	0.56	0.59	0.75	0.73	0.74	0.82	0.82	0.82	0.82

(b) Transformer and LLM models

Model	Non-slang			Slang			Avg. macro			Avg. weighted			Acc
	Pr	Rec	F1	Pr	Rec	F1	Pr_M	Rec_M	F1_M	Pr_W	Rec_W	F1_W	
BERT-L (Uncased)	0.90	0.94	0.92	0.76	0.63	0.69	0.83	0.79	**0.80**	0.86	0.87	0.87	**0.87**
RoBERTa-L	0.88	0.91	0.90	0.66	0.58	0.62	0.77	0.75	0.76	0.84	0.84	0.83	0.84
XLNet-L	0.87	0.95	0.90	0.60	0.49	0.54	0.73	0.72	0.72	0.83	0.82	0.82	0.82
ALBERT-XXL	0.89	0.94	0.92	0.76	0.61	0.68	0.82	0.78	0.80	0.86	0.86	0.85	0.86
GPT-4o (ZS)	0.94	0.79	0.86	0.54	0.82	0.65	0.74	0.80	0.75	0.80	0.80	0.80	0.80
GPT-4o (FS)	0.94	0.73	0.82	0.48	0.84	0.61	0.71	0.78	0.71	0.75	0.75	0.75	0.75
GPT-4o-Mini (FT)	0.91	0.92	0.91	0.72	0.66	0.69	0.81	0.79	**0.80**	0.86	0.86	0.86	**0.86**
GPT-4o (FT)	0.85	0.90	0.89	0.76	0.63	0.69	0.81	0.77	0.79	0.85	0.85	0.85	0.85
LLaMA-3.1-8B (FT)	0.37	0.62	0.47	0.86	0.62	0.72	0.61	0.62	0.59	0.75	0.67	0.53	0.67
LLaMA-3.1-70B (FT)	0.87	0.64	0.73	0.36	0.69	0.47	0.61	0.66	0.60	0.75	0.65	0.68	0.65
Models with Sentiment & Emotion													
BERT-L +S	0.90	0.95	0.92	0.79	0.64	0.76	0.84	0.80	0.84	0.88	0.88	0.88	0.88
BERT-L +E	0.90	0.94	0.92	0.78	0.63	0.76	0.84	0.79	0.84	0.87	0.87	0.87	0.87
BERT-L +S+E	0.90	0.95	0.92	0.81	0.65	0.78	0.86	0.80	**0.85**	0.89	0.89	0.89	**0.89**
ALBERT +S	0.90	0.93	0.92	0.77	0.66	0.74	0.83	0.80	0.83	0.87	0.87	0.87	0.87
ALBERT +E	0.90	0.92	0.91	0.76	0.66	0.73	0.83	0.79	0.82	0.87	0.87	0.86	0.87
ALBERT +S+E	0.91	0.92	0.91	0.72	0.69	0.77	0.82	0.81	0.84	0.88	0.88	0.88	0.88

Table 3: Model performance on the SlangTrack test set. Panel (a): deep learning models. Panel (b): transformer and large language models, including ablations with sentiment (S) and emotion (E) features. ZS = Zero-shot, FS = Few-shot, FT = Fine-tuned, S = Sentiment, E = Emotion.

detection, achieving a precision of 0.36 and a recall of 0.69. This indicates that without targeted fine-tuning, LLaMA models are less suited to the fine-grained distinctions required for slang disambiguation.

A per-word breakdown of BERT-large-uncased (Table 5) reveals substantial variation across target words. Items such as *climber* and *salty* achieve high macro-F1, while more ambiguous or infrequent slang senses (e.g., *cucumber*, *mammy*) yield lower slang-F1. This variation reflects differences in sense ambiguity and class balance, underscoring the need for per-word analysis beyond aggregate performance metrics.

Transformer fine-tuning produced the strongest performance, and adding sentiment and emotion features yielded small but consistent improvements. For BERT-large and ALBERT-xxlarge, affective cues increased macro-F1 by approximately 0.03–0.05 and weighted F1 by around 0.01–0.02, primarily through slightly higher recall on slang instances. This suggests that affective information helps distinguish some figurative or evaluative uses from literal ones. Overall, affective features provide modest, model-specific gains and function as a supplementary rather than primary signal. Variance across runs is reported in Appendix 14. As shown in Table 10, improvements from sentiment and emotion features exceed the observed variance. Statistical significance is confirmed via paired bootstrap testing (Appendix 15).

7 Error Analyses

We analysed 100 misclassified instances from our top-performing slang detection model (BERT-L + Sentiment + Emotion), categorising them into types (Table 12).

Bad neighbours (23%): Misclassification often

Target Words	Example Sentences	Prediction Before Sentiment & Emotion	→	Sentiment	Emotion	Prediction After Sentiment & Emotion
Mammy	First 'black' woman to win an Oscar played **a mammy** slave. .. last one to win it also played a slave. Some see this as progress.	Non-Slang	→	Negative	Sadness	Slang
Germ	My room mate won't stop calling me a walking **germ** factory just because I have a cold.	Non-Slang	→	Negative	Anger	Slang
Chronic	North Korea is known for its **chronic** secrecy and isolation. But in recent years, despite its **chronic** struggles, there's been a noticeable rise in underground markets.	Slang	→	Neutral	Fear	Non-Slang
Eat	Just finished that brutal workout can't wait to **eat!** My abs are on fire, but hey, no pain, no gain, right?	Slang	→	Positive	Joy	Non-Slang

Table 4: Impact of sentiment and emotion on slang classification. The table shows examples of *Prediction before sentiment & emotion* and their corrected outputs using *Prediction after sentiment & emotion*.

Word	Cnt	F1 Non-slang	F1 Slang	Macro F1
Bmw	1,097	0.995	0.421	0.708
Brownie	964	0.872	0.800	0.836
Chronic	1,685	0.931	0.598	0.765
Climber	642	0.964	0.832	0.898
Cucumber	1,051	0.960	0.415	0.688
Eat	3,032	0.905	0.454	0.680
Germ	832	0.848	0.632	0.740
Mammy	1,070	0.929	0.458	0.693
Rodent	1,073	0.844	0.651	0.748
Salty	1,266	0.870	0.909	0.890

Table 5: Per-word F1 scores for BERT-large-uncased.

arises from the presence of nearby words that introduce strong pragmatic signals, such as abusive language, drug references, or harsh tones. Although such cues are not explicitly modelled in SlangTrack, they may co-occur with slang usage in natural data and implicitly influence model predictions. Importantly, SlangTrack does not aim to model toxicity or harmful intent; rather, these signals are treated as incidental contextual factors that can skew interpretation.

Proper nouns (10%): Proper nouns, especially those appearing as bi-grams or tri-grams, can confuse the model due to their compact and informal structure. These structures may be misinterpreted as slang or colloquial expressions, particularly when they lack distinguishing features.

Lost in length (13%): Very long instances with multiple clauses, or extremely short instances with limited contextual cues, can challenge the model. In such cases, important pragmatic signals may be diluted, truncated, or insufficiently represented in the input, reducing the model's ability to distinguish slang from literal usage.

Polysemy (17%): Polysemy refers to target words with multiple meanings, both slang and non-slang, leading to misclassification when the immediate context is insufficient to disambiguate. For example, the word "germ" can mean a microorganism or serve as a slang insult. Polysemy concerns the target word's inherent multiple meanings rather than the surrounding context.

Ambiguity (7%): Ambiguity arises when the broader sentence context creates uncertainty, even if the target word's meaning is clear in isolation. For instance, "salty" can mean "bitter" slang or "overly seasoned" literal. The word itself has a clear meaning in both cases, but without strong contextual cues, the model struggles to determine which sense is intended. Ambiguity is about the sentence creating uncertainty rather than the word having multiple inherent meanings.

Unknown (30%): Unconventional abbreviations, rare slang, or novel expressions pose challenges. Such terms often deviate from standard language patterns, making classification difficult, especially if they are absent from the model's training data.

8 Conclusion

We address the task of slang detection by focusing on words that can be used in both slang and non-slang senses, an aspect that has received limited attention in prior work. This distinction enables a more fine-grained analysis of lexeme-specific and pragmatically driven meaning alternations, beyond aggregate classification accuracy. We introduce SlangTrack, a corpus of dual-meaning words, and benchmark a range of neural architectures, fine-tuned transformer-based language models, and large language models. Our results show that fine-tuned transformer models, particularly BERT-large-uncased, achieve the strongest performance on this task, while incorporating sentiment and emotion features yields modest but consistent improvements in ambiguous cases. SlangTrack thus provides a controlled benchmark for studying contextual slang usage across different lexical items.

9 Limitations

Our dataset integrates examples from heterogeneous sources that differ in register and historical period. CCOHA reflects earlier formal written English, while Twitter provides contemporary informal. These source differences may affect recall for slang senses that vary across time or context, but they also reflect real lexical change and support the study of sense variation across registers. Because CCOHA and Twitter differ substantially in corpus size and temporal span, we do not interpret raw frequency differences between them as direct evidence of diachronic slang trends; instead, our analyses focus on sense-level annotation and statistically controlled comparisons of slang versus non-slang usage across corpora. Although SlangTrack includes temporal metadata, the present study does not implement time-aware modelling, leaving explicit modelling of continuous semantic trajectories and slang adoption dynamics to future work.

While SlangTrack captures instance-internal pragmatic cues (e.g., affective stance and figurativity) expressed locally within individual textual instances, it does not explicitly model discourse-level pragmatics such as speaker alignment or conversational implicature spanning multiple interactions. The task therefore models local pragmatic interpretation rather than full discourse-level slang understanding. As with naturally occurring language data, slang and non-slang usages are not perfectly balanced across target words. This natural sparsity is addressed through macro-level evaluation and per-word analyses. Another limitation is the narrow lexical coverage (10 target words), adopted as a controlled pilot setting. Rather than maximising absolute performance, the results provide diagnostic insights into lexeme-dependent slang detection and motivate future work on broader coverage and richer contextual modelling.

References

Michael Adams. 2012. *Slang: The people's poetry*. Oxford University Press.

Reem Alatrash, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte Im Walde. 2020. CCOHA: Clean corpus of historical American English. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6958–6966.

Michele Bevilacqua and Roberto Navigli. 2020. Breaking through the 80% glass ceiling: Raising the state of the art in word sense disambiguation by incorporating knowledge graph information. In *Proceedings of the conference-Association for Computational Linguistics. Meeting*, pages 2854–2864. Association for Computational Linguistics.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of african-american english. *arXiv preprint arXiv:1608.08868*.

Julie Coleman. 2012. *The life of slang*. Oxford University Press, USA.

Jader Martins Camboim de Sá, Marcos Da Silveira, and Cédric Pruski. 2024. Survey in characterization of semantic change. *arXiv preprint arXiv:2402.19088*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Shehzaad Dhuliawala, Diptesh Kanojia, and Pushpak Bhattacharyya. 2016. SlangNet: A WordNet like resource for English slang. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4329–4332.

Bethany K Dumas and Jonathan Lighter. 1978. Is slang a word for linguists? *American Speech*, 53(1):5–17.

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 359–369.

Jonathon Green. 2015. *The vulgar tongue: Green's history of slang*. Oxford University Press, USA.

Anshita Gupta, Sanya Bathla Taneja, Garima Malik, Sonakshi Vij, Devendra K Tayal, and Amita Jain. 2019. SLANGZY: A fuzzy logic-based algorithm for English slang meaning Selection. *Progress in Artificial Intelligence*, 8:111–121.

Janosch Haber and Massimo Poesio. 2024. Polysemy—evidence from linguistics, behavioral science, and contextualized language models. *Computational Linguistics*, 50(1):351–417.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv:1605.09096*.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. GlossBERT: BERT for word sense disambiguation with gloss knowledge. *arXiv preprint arXiv:1908.07245*.

Daphna Keidar, Andreas Opedal, Zhijing Jin, and Mrinmaya Sachan. 2022. Slangvolution: A causal analysis of semantic change and frequency dynamics in slang. *arXiv preprint arXiv:2203.04651*.

Chloe Kiddon and Yuriy Brun. 2011. That's what she said: double entendre identification. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 89–94.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 388–395.

Sawan Kumar, Sharmistha Jat, Karan Saxena, and Partha Talukdar. 2019. Zero-shot word sense disambiguation using sense definition embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5670–5681.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. *arXiv preprint arXiv:1806.03537*.

Z Lan. 2019. Albert: A LiteBert for Self-supervised Learning of Language Representations. *arXiv preprint arXiv:1909.11942*.

Yann LeCun and Yoshua Bengio. 1995. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995.

Gang Liu and Jiabao Guo. 2019. Bidirectional LSTM with attention mechanism and convolutional layer for text classification. *Neurocomputing*, 337:325–338.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Li Lucy and David Bamman. 2021. Style variation and social meaning in online communities. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4745–4760. Association for Computational Linguistics.

Theo Lynn, Patricia Takako Endo, Pierangelo Rosati, Ivanovitch Silva, Guto Leoni Santos, and Debbie Ging. 2019. A comparison of machine learning approaches for detecting misogynistic speech in urban dictionary. In *2019 International Conference on Cyber Situational Awareness, Data Analytics And Assessment (Cyber SA)*, pages 1–8. IEEE.

George A. Miller. 1995. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11):39–41.

Ke Ni and William Yang Wang. 2017. Learning to explain non-standard English words and phrases. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 413–417, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Alok Ranjan Pal and Diganta Saha. 2015. Detection of slang words in e-data using semi-supervised learning. *arXiv preprint arXiv:1702.04241*.

Zhengqi Pei, Zhewei Sun, and Yang Xu. 2019. Slang detection and identification. In *Proceedings of the 23rd conference on computational natural language learning (CoNLL)*, pages 881–889.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Francesco Periti, Pierluigi Cassotti, Haim Dubossarsky, and Nina Tahmasebi. 2024. Analyzing semantic change through lexical replacements. *arXiv preprint arXiv:2404.18570*.

Rhys J Sandow, George Bailey, and Natalie Braber. 2024. Language change is wicked: semantic and social meaning of a polysemous adjective. *English Language & Linguistics*, 28(1):135–156.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23. International Committee for Computational Linguistics.

Zhewei Sun, Qian Hu, Rahul Gupta, Richard Zemel, and Yang Xu. 2024. Toward informal language processing: Knowledge of slang in large language models. *arXiv preprint arXiv:2404.02323*.

Zhewei Sun, Richard Zemel, and Yang Xu. 2021. A computational framework for slang generation. *Transactions of the Association for Computational Linguistics*, 9:462–478.

Zhewei Sun, Richard Zemel, and Yang Xu. 2022. Semantically informed slang interpretation. *arXiv preprint arXiv:2205.00616*.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2021. Survey of computational approaches to lexical semantic change detection. *Computational approaches to semantic change*, 6(1).

Liang Wu, Fred Morstatter, and Huan Liu. 2018. Slangsd: building, expanding and using a sentiment dictionary of slang words for short-text sentiment classification. *Language Resources and Evaluation*, 52:839–852.

Ifeoluwa Wuraola, Nina Dethlefs, and Daniel Marciniak. 2024. Understanding slang with LLMs: Modelling

cross-cultural nuances through paraphrasing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15525–15531, Miami, Florida, USA. Association for Computational Linguistics.

Zhilin Yang. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

10 Implementation Details

Hardware Configuration

Deep Learning Models. All CNN and BiLSTM experiments were conducted on a single Tesla V100 GPU with 32 GB RAM using the Google Colab Pro+ platform.[17]

Fine-Tuning Experiments (LMs and LLMs). Fine-tuning for transformer-based language models and large language models (LLMs) was performed on a single NVIDIA A100 GPU (80 GB RAM).

Software Frameworks

All experiments used Python 3.10.12. Evaluation metrics were computed using scikit-learn.[18] Transformer models were trained with three random seeds; performance variation was minimal (macro-F1 standard deviation 0.01).

Deep Learning Model Frameworks

TensorFlow[19] and Keras[20] were used for implementing CNN and BiLSTM architectures. Hyperparameter tuning was performed using Keras Tuner.[21] scikit-learn utilities were used to compute precision, recall, and F1-score.

Transformer Model Frameworks

SimpleTransformers[22] and Hugging Face Transformers[23] enabled loading and fine-tuning of models such as BERT. Stratified K-fold cross-validation was implemented with scikit-learn.

Deep Learning (CNN and BiLSTM) Architecture Setup

Embedding Resources. FastText embeddings (`wiki-news-300d-1M.vec`), trained on Wikipedia 2017, UMBC WebBase, and statmt.org news (16B tokens), were used.[24] GloVe embeddings (300d, Common Crawl, 840B tokens) were loaded. BERT embeddings from `bert-base-uncased` were extracted for use in CNN and BiLSTM models.

Tokenisation and Sequence Processing. Sentences were tokenised using the Keras `Tokenizer`. Sequences were padded to a fixed length (95th percentile of training sentence lengths) using `pad_sequences`.

Hyperparameter Tuning. CNN and BiLSTM configurations were optimised via Keras Tuner's Random Search. Fifty trials were run per architecture, tuning embedding dimensions, hidden sizes, dropout, learning rate, and optimiser. Selected hyperparameters appear in Table 13.

Fine-Tuning Procedures for GPT-4o and LLaMA Models

Training Setup and Specifications. GPT-4o, GPT-4o-mini, LLaMA-3.1-70B-Instruct-Reference, and LLaMA-3.1-8B-Instruct were fine-tuned using the same training and validation splits as BERT models. GPT-4o models were fine-tuned via OpenAI's API. LLaMA models were fine-tuned using Together AI's API,[25], incorporating LoRA for memory-efficient training. Each model was trained to classify sentences as `slang` or `non-slang` using structured JSONL prompt–completion pairs.

Structured Prompt Format. A consistent prompt template was used for GPT-4o and LLaMA models: `Your task is to classify the sentence as either 'slang' or 'non-slang.' Please respond only with 'slang' or 'non-slang.'`

11 Exploratory ANOVA Analysis Between Slang and Non-Slang Texts

To provide a descriptive view of stylistic differences between slang and non-slang instances, we conducted an exploratory one-way ANOVA using automatically derived sentiment and emotion scores.[26]

[17] https://colab.research.google.com/
[18] https://scikit-learn.org/stable/
[19] https://www.tensorflow.org/
[20] https://keras.io/
[21] https://keras-team.github.io/keras-tuner/
[22] https://simpletransformers.ai
[23] https://huggingface.co/transformers/
[24] https://fasttext.cc/docs/en/english-vectors.html
[25] https://together.ai
[26] All affective scores are produced by an external classifier and were not manually annotated; results should therefore be interpreted only as descriptive trends rather than causal linguistic evidence.

11.1 Sentiment Differences

Figure 1 shows the distribution of positive, neutral, and negative sentiment. ANOVA results (Table 6) indicate statistically detectable but modest differences between classes. Slang instances tend to be slightly more negative, while non-slang instances are more often neutral.

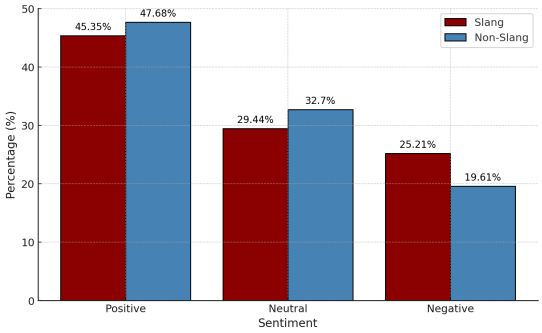

Figure 1: Sentiment distribution across slang and non-slang instances.

Sentiment	F	p	Significance
Positive	45.78	2.1×10^{-10}	Significant
Negative	38.54	4.3×10^{-8}	Significant
Neutral	12.92	3.0×10^{-4}	Significant

Table 6: Exploratory ANOVA results for sentiment.

11.2 Emotion Differences

Emotion scores (Figure 2) show detectable differences for a subset of categories (Table 7). Slang instances exhibit slightly higher anger and sadness, while other emotions show no meaningful differences.

Emotion	F	p	Significance
Anger	9.61	1.9×10^{-3}	Significant
Sadness	5.35	3.7×10^{-2}	Significant
Fear	0.11	0.74	Not Significant
Joy	0.32	0.57	Not Significant
Love	1.52	0.22	Not Significant
Surprise	0.03	0.85	Not Significant

Table 7: Exploratory ANOVA results for emotion categories.

12 Appendix: Full Sense Inventories for SlangTrack Target Words

Table 8 lists all attested senses for each target word in the SlangTrack dataset. Sense IDs (S1–Sn) correspond to the annotation scheme used throughout the paper. This table enumerates the sense inventories only.

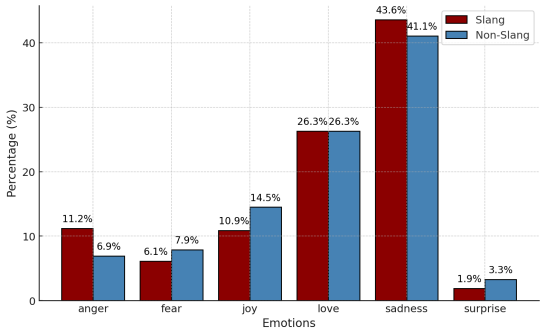

Figure 2: Emotion score distribution across slang and non-slang instances.

13 Statistical Analysis of Diachronic Shifts in Slang vs. non-slang Sense Usage

We evaluate whether the relative frequency of pre-annotated slang versus non-slang senses differs between COHA (historical data) and Twitter (contemporary data). Each annotated token represents a binary outcome (slang or non-slang), enabling comparison of proportions across periods using the standard two-proportion z-test. This analysis examines redistribution in the usage of existing senses rather than the emergence of new ones.

13.1 Two-Proportion z-Test

For a given word w, let $p_{\text{COHA}}(w)$ and $p_{\text{Twitter}}(w)$ denote the observed proportions of slang-labelled tokens in COHA and Twitter. The null hypothesis states that the true slang probabilities are equal:

$$H_0 : p_{\text{COHA}}(w) = p_{\text{Twitter}}(w),$$

$$\mathrm{H}_1 : p_{\text{COHA}}(w) \neq p_{\text{Twitter}}(w).$$

The pooled estimate under H_0 is:

$$\hat{p} = \frac{s_{\text{COHA}} + s_{\text{Twitter}}}{n_{\text{COHA}} + n_{\text{Twitter}}},$$

where s_{COHA} and s_{Twitter} are slang counts, and $n_{\text{COHA}}, n_{\text{Twitter}}$ are total token counts in each corpus. The standard error of the difference is:

$$SE = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_{\text{COHA}}} + \frac{1}{n_{\text{Twitter}}}\right)}.$$

The test statistic is:

$$z = \frac{\hat{p}_{\text{Twitter}} - \hat{p}_{\text{COHA}}}{SE}.$$

Word	Sense Inventory (S1–Sn)
Eat	S1: Consume food. S2: Perform oral sex on a woman. S3: Make money or absorb a financial loss ("take the loss"). S4: Rob someone in a low-violence street context. S5: Defeat, destroy, or overwhelm. S6: Irritate or annoy.
BMW	S1: German automobile brand. S2: Derogatory reference to Black individuals ("Black Man Working," etc.). S3: Acronym for "Be My Wife."
Brownie	S1: Chocolate dessert. S2: Silly or foolish person. S3: Racially offensive reference to brown-skinned individuals. S4: Marijuana-infused edible. S5: Vulgar reference to the anus. S6: Unit of social credit ("brownie points"). S7: Personal name or nickname. S8: Frequent collocate used in fixed expressions.
Chronic	S1: High-quality cannabis. S2: Medical term for a long-lasting or persistent condition. S3: Excellent or outstanding. S4: Negative extreme (e.g., a "chronic" habit or person). S5: Proper name (album, song, title).
Climber	S1: Person attempting upward social mobility. S2: Rock climber (sport). S3: Climbing plant. S4: Burglar who gains access by climbing.
Germ	S1: Microorganism. S2: Cigarette slang used in institutional settings. S3: Offensive term for people of German descent. S4: Contemptible or unpleasant person. S5: Proper name (company or brand). S6: Frequent collocate in compounds.
Mammy	S1: Racialised stereotype of a Black woman domestic worker. S2: Dialectal or literal "mother." S3: Slang for a large amount (e.g., "money's mammy"). S4: Proper name (brand, company, or artistic title).
Cucumber	S1: Penis (slang). S2: Vegetable. S3: Proper name (company, brand, or software).
Rodent	S1: Insult for someone unattractive, untrustworthy, or unintelligent. S2: Animal of the order *Rodentia*.
Salty	S1: Irritated, annoyed, or resentful. S2: Tasting of salt. S3: Old, worn, or well-used. S4: Tough, aggressive, or hardened. S5: Crude, obscene, or vulgar. S6: Proper name (animal, brand, or song title).

Table 8: Complete sense inventories for all SlangTrack target words. Slang and non-slang senses are shown together. Sense IDs (S1–Sn) correspond to those used in the annotation guidelines.

Two-sided p-values are obtained from the standard normal distribution.

13.2 Worked Example: *Mammy*

For *Mammy*, the annotated sense inventory includes both slang and non-slang meanings. Slang senses: (1) stereotyped Black woman ("mammy" caricature), (3) figurative abundance sense (e.g., "money's mammy"). Non-slang senses: (2) literal 'mother', (4) proper-name, company, or song title use.

COHA. Slang tokens: 97. Total tokens: 194.

$$\hat{p}_{\text{COHA}} = 97/194 = 0.500.$$

Twitter. Slang tokens: 69. Total tokens: 876.

$$\hat{p}_{\text{Twitter}} = 69/876 \approx 0.0788.$$

Pooled estimate.

$$\hat{p} = \frac{97 + 69}{194 + 876} = \frac{166}{1070} \approx 0.1551.$$

Standard error.

$$SE \approx 0.0287.$$

Test statistic.

$$z \approx -14.66, \qquad p < 0.0001.$$

This indicates a substantial decrease in the slang usage of *Mammy* from COHA to Twitter, with the racialised sense becoming rare in contemporary data.

13.3 Results for All Target Words

Table 9 reports slang proportions for COHA and Twitter, their differences, and z-test significance levels.

13.4 Interpretation

Three items show statistically significant redistribution in slang usage between COHA and Twitter: *Salty* (increase), *Mammy* (decrease), and *Chronic* (decrease). For all other words, changes are small

Word	p_{COHA}	p_{Twitter}	Δp	z	p	Sig.
Eat	0.100	0.120	+0.020	1.67	0.0960	-
BMW	0.022	0.012	-0.010	-1.16	0.2450	-
Brownie	0.283	0.265	-0.018	-0.45	0.6550	-
Chronic	0.280	0.225	-0.055	-2.59	0.0097	**
Climber	0.193	0.219	+0.026	0.65	0.5150	-
Germ	0.110	0.091	-0.019	-0.78	0.4370	-
Mammy	0.500	0.079	-0.421	-14.66	<0.0001	***
Cucumber	0.057	0.072	+0.015	0.69	0.4890	-
Rodent	0.292	0.309	+0.017	0.46	0.6470	-
Salty	0.193	0.742	+0.549	18.12	<0.0001	***

Table 9: Slang usage proportions for COHA vs. Twitter. Significance: $*p < .01$, $**p < .001$ (two-sided).

and not significant. The results illustrate diachronic variation in the distribution of attested senses, consistent with usage-based accounts of sense competition and shifting pragmatic salience.

14 Variance Estimation

To provide a transparent account of performance variability, we report macro-F1 scores together with variance estimates for all models (see Table 10). Transformer models were fine-tuned using three distinct random seeds, and we therefore report macro-F1 $\pm$ standard deviation across these runs. For deep learning models, which are deterministic under our setup, variance was estimated using 1,000-sample non-parametric bootstrap resampling over the test predictions.

Model	Non-slang F1	Slang F1	Macro-F1 $\pm$ SD
Deep Learning Models			
BiLSTM (BERT)	0.91	0.58	0.75 ± 0.012
BiLSTM (GloVe)	0.90	0.62	0.76 ± 0.015
BiLSTM (FastText)	0.90	0.62	0.76 ± 0.014
CNN (BERT)	0.90	0.60	0.75 ± 0.017
CNN (GloVe)	0.87	0.61	0.74 ± 0.013
CNN (FastText)	0.90	0.62	0.76 ± 0.012
BiLSTM-CRF (full features)	0.86	0.48	0.67 ± 0.020
CNN-CRF (full features)	0.89	0.59	0.74 ± 0.019
Transformer Models			
BERT-large-uncased	0.92	0.69	0.80 ± 0.010
RoBERTa-large	0.90	0.62	0.76 ± 0.011
XLNet-large	0.90	0.54	0.72 ± 0.012
ALBERT-xxlarge-v2	0.92	0.68	0.80 ± 0.010
Models with Sentiment & Emotion			
BERT-L +S	0.92	0.76	0.84 ± 0.010
BERT-L +E	0.92	0.76	0.84 ± 0.011
BERT-L +S+E	0.92	0.78	$\mathbf{0.85} \pm 0.012$
ALBERT +S	0.92	0.74	0.83 ± 0.011
ALBERT +E	0.91	0.73	0.82 ± 0.012
ALBERT +S+E	0.91	0.77	0.84 ± 0.013

Table 10: Performance of deep learning and transformer models on the SlangTrack test set, including models enriched with sentiment (S) and emotion (E) features.

15 Statistical Significance of Sentiment and Emotion Features

To verify that the performance improvements from sentiment and emotion features are statistically meaningful, we conducted paired significance testing on the transformer models.

Model	Comparison	Macro-F1	p-value
BERT-large-uncased	Base $\rightarrow$ +S	+0.04	0.012
	Base $\rightarrow$ +E	+0.04	0.018
	Base $\rightarrow$ +S+E	+0.05	0.003
ALBERT-xxlarge-v2	Base $\rightarrow$ +S	+0.03	0.021
	Base $\rightarrow$ +E	+0.02	0.025
	Base $\rightarrow$ +S+E	+0.04	0.009

Table 11: Paired bootstrap significance testing for sentiment and emotion features. Each block shows the comparisons between the base model and variants with sentiment (S), emotion (E), or both.

Test procedure. For each transformer-based model, we compare the base fine-tuned model with its affect-enhanced variants using paired bootstrap resampling (Koehn, 2004). Resampling is performed over test predictions from a fixed fine-tuning configuration, using identical test instances for both models. We draw 10,000 bootstrap samples with replacement and compute the difference in macro-F1 for each sample, yielding an empirical distribution of performance differences. Two-sided p-values are estimated as the proportion of samples in which the difference includes zero.

Results. Table 11 reports the mean macro-F1 improvements (ΔMacro-F1) and associated p-values. For both BERT-large and ALBERT-xxlarge, the combined sentiment and emotion configuration (+S+E) yields statistically significant improvements ($p < 0.05$). Individual sentiment-only (+S) and emotion-only (+E) features produce smaller gains that are not consistently significant across models. These findings are consistent with the low run-to-run variance observed across random seeds (Appendix 14), indicating that the improvements are robust to sampling variation rather than artefacts of random initialisation.

16 Annotation Challenges in Historical Corpus Data

These examples demonstrate why annotating words that have both slang and non–slang senses is cognitively demanding, particularly in long, descriptive, historical corpora such as COHA. One of our target words, *salty*, can express a wide range of meanings, and the annotator must determine which sense is intended based on extended narrative context. These examples illustrate how COHA's long, descriptive, and multi-layered passages often involving figurative language, shifting narrative perspectives, and domain-specific idioms significantly increase the

complexity of sense annotation. As a result, annotations drawn from historical corpora require more time, attention, and contextual reasoning than annotations drawn from short, contemporary, or informal texts.

Sense Inventory for *salty*

- **S1: Irritated, annoyed, or resentful** slang sense referring to emotional upset or reactive behaviour.
- **S2: Tasting of salt** — literal reference to flavour, salinity, or physical salt content.
- **S3: Old, worn, or well-used** describes objects or individuals that appear aged or weathered.
- **S4: Tough, aggressive, or hardened** slang sense used for rugged, experienced, or hardened individuals.
- **S5: Crude, obscene, or vulgar** slang sense used for coarse or inappropriate language or behaviour.
- **S6: Proper name** used when *salty* appears as a title or name (animal, brand, song title).

Example 1 (COHA T1)

My feelings are hurt, and no one loves me. It doesn't matter what I look like. A large bowl of snack food or a box of chocolates is so 'good' that it disappears. (e) "I am just as good, strong, and brave as they are." If I eat heart, kidneys, and liver, then I will be strong like a lion. (f) "I almost starved to death during period X." People who have been deprived of food continuously often will eat whatever they can, including rich and salty snack foods @ @ @ @ @ @ @ @ @ @ d. Due to the following aspects: (a) to make it less extreme, when a person who habitually consumes a high-fat diet must reduce fat intake to a moderate level; (b) to limit or restrict the meaning of and suggests a difference that limits, restricts, or adapts to a new purpose, as when one who has habitually consumed a high-fat, rich diet must follow a low-fat, low-calorie diet until sufficient weight is lost that the gall bladder can be removed; (c) to make a minor change in, as a patient with an ulcer. Such dietary adjustments are often recommended gradually rather than abruptly. Medical guidance typically emphasizes moderation rather than total elimination. Patients are advised to monitor portion size and ingredient content carefully. Attention to fat and salt intake is presented as part of long-term health management.

Sense-Identification Notes:

This passage blends nutritional explanation, emotional examples, and medical reasoning, making the annotator process several shifts in meaning before encountering the literal use of salty. The long, layered structure increases the cognitive load and requires careful reading. The intended sense is a literal reference to salty food.

Label: *S2 Non-slang*

Example 2 (COHA T1)

Back into the hole you crawled @ @ @ @ @ @ @ @ @ and get out of here. Hey, hey, hey, hey, hey. Hey, what're you doing? Pretty hot out here. Must be what, 90? It's gonna get real toasty in there. Hey, now, now, let's not get salty here — no need to get angry. I'll tell you what. You give everything you got in the register, forget about pumping gas, and we'll be on our merry way. The purse, too. Please, go away now. No! No, don't! No! No, don't! (SHOUTING) Whoa, whoa! Could you hold it there? MAN: Ging, guy. Baker, what is this? That's, um, caviar. You know, that stuff comes all the @ @ @ @ @ @ @ @ @ — you kidding me? I get a stupid little sea wrapped pound cake with a candle on it for my birthday, and this guy's spending $20 an ounce on this stuff? If you had the stars, it would be a whole different war. Yeah. How do you eat it? Please put it on one of those crackers, Danny don't get salty about it. RUIZ: Oh, that's disgusting. Danny, Danny, man, I got to clean that stuff up... Hey, keep your voice down, all right? You don't need to snap at everyone in the room. Everybody's already on edge, and you're making it worse. Take a breath and calm yourself down. This doesn't have to turn ugly if you don't let it. Now move, before someone really loses their temper.

Sense-Identification Notes:

The expression "let's not get salty here" appears during an escalating confrontation marked by shouting, threats, and attempts to control anger. In this context, salty clearly refers to emotional irritation or annoyance rather than taste or physical properties. The surrounding dialogue explicitly references anger management ("no need to get angry," "calm yourself down," "loses their temper"), which firmly supports interpretation as S1.

Label: *S1 Slang*

Example 3 (COHA T2)

Third Son ends up with the princess and half the kingdom. He fitted right in. Only Emily knew he didn't belong, and it gave a kind of edge to his performance, she felt. She ran him through his lines herself. (Leon played the older two sons.) She put an extra, salty twang in his voice. The real Third Son, meanwhile more handsome, with less character – lay face-up backstage, grinning vacantly. Emily had never actually planned to be @ @ @ @ @ @ @ @ @ @, thought of it as temporary work. "They would taste now; it's hot in here," she told him. @ @ @ @ @ @ @ @ @ his jaw, a tiny muscle into which he poured all the concentrated tension of his body. "You invited me here, lady..." He stood straighter, setting his shoulders as if bracing for impact. The words came out sharper than before, clipped and deliberate. Emily nodded, encouraging him to hold that tone. The edge made the character sound older, tougher, less naïve. It was no longer a boy pleading, but a man demanding. Even from backstage, the difference in force was audible. The performance carried a harder, more aggressive energy. The added roughness gave the scene its weight.

Sense-Identification Notes:

In this theatrical context, salty modifies twang to describe a deliberately sharpened and hardened vocal quality. The surrounding performance-related cues (e.g., tension, force, and aggression in delivery) indicate a stylistic choice rather than a literal taste or emotional irritation. This usage aligns with the slang sense denoting toughness or hardness.

Label: *S4 Slang*

Error Category	Examples	Reason for Misclassification	Gold Label → Predicted Label
Bad Neighbors	I think y'all understand the intense hate and fear for that **rodent**-looking motherfucker.	The strong slang word "motherfucker" triggers misclassification, while "rodent" adds a negative tone but isn't slang.	Non-Slang → Slang
Proper Nouns	Wow, believe still remember **brownie** smile song girl scout memories. Good burger, man. I wish you could come to the sweet brownie party.	Phrases like "Brownie Smile Song" and "sweet brownie party" are proper nouns. Informal phrasing misleads the model into treating them as slang.	Non-Slang → Slang
Lost in Length	Post-1960s growth, a small, expensive underclass resulted in structural problems... **chronic** joblessness and welfare dependency.	Long, complex sentences make identifying context difficult, and pre-processing can reduce clarity.	Non-Slang → Slang
Polysemy	@Officer_Grayson Once **a germ**, always **a germ**. He's as unclean as pork.	The word "germ" has multiple meanings, either as a microorganism or an insult. Lack of clear context causes errors.	Non-Slang → Slang
Polysemy	As a tiny **rodent**.. I see things from a unique angle. Like that guy over there... he's not wearing underpants.	The metaphor "tiny rodent" was interpreted literally instead of as slang, leading to misclassification.	Slang → Non-Slang
Ambiguity	The book's protagonist is **a mammy** figure who is both nurturing and deeply flawed, and becomes a symbol of resistance against systemic oppression.	The context links "mammy" to literary analysis, suggesting non-slang usage, but informal or stereotypical connotations mislead the model.	Non-Slang → Slang
Ambiguity	My mom is really starting to get on my fucken nerves being the **germ** freak she is.	The word "germ" can be literal (bacteria) or slang (obsession with cleanliness). Ambiguous context caused misclassification.	Slang → Non-Slang
Unknown	Ugh, can't wait to **eat** something after this workout! Abs are killing me, lol hoebag move, though.	Informal abbreviations like "lol" and rare slang terms like "hoebag" confuse the model.	Non-Slang → Slang

Table 12: Examples of misclassified samples for each error category.

Deep Learning, Transformer, and Large Language Models	
Model	**Parameters**
Fine-Tuned Transformer Models (BERT, RoBERTa, ALBERT, XLNet)	`num train epochs = 30, learning rate = 4e-5, train batch size = 64, eval batch size = 64, Optimiser = AdamW`
Fine-Tuned BERT and ALBERT with Sentiment and Emotion Analysis	`num train epochs = 30, learning rate = 4e-5, train batch size = 64, eval batch size = 64, Optimiser = AdamW`
BiLSTM + GloVe embeddings	`Embedding Dimension = 300, BiLSTM Units = 32, Dense Units = 64, Dropout Rate = 0.2, Optimiser = Adam, Learning Rate = 2.93e-03, epochs = 30`
BiLSTM + FastText embeddings	`Embedding Dimension = 300, BiLSTM Units = 256, Dense Units = 64, Dropout Rate = 0.3, Optimiser = Adam, Learning Rate = 7.02e-04, epochs = 30`
BiLSTM + BERT embeddings	`Embedding Dimension = 768, BiLSTM Units = 128, Dense Units = 128, Dropout Rate = 0.3, Optimiser = RMSprop, Learning Rate = 7.44e-03, epochs = 30`
CNN + GloVe embeddings	`Embedding Dimension = 300, Conv Units = 128, Dense Units = 32, Dropout Rate = 0.4, Optimiser = Adam, Learning Rate = 1.00e-03, epochs = 30`
CNN + FastText embeddings	`Embedding Dimension = 300, Conv Units = 224, Dense Units = 128, Dropout Rate = 0.2, Optimiser = RMSprop, Learning Rate = 1e-03, epochs = 30`
CNN + BERT embeddings	`Embedding Dimension = 768, Conv Units = 64, Dense Units = 32, Dropout Rate = 0.3, Optimiser = RMSprop, Learning Rate = 1e-03, epochs = 30`
Fine-Tuned LLaMA Models (LLaMA-3.1-70B-Instruct, LLaMA-3.1-8B-Instruct)	`num train epochs = 30, learning rate = 1e-5, train batch size = 64, LoRA Enabled = True, LoRA Rank = 64, LoRA Alpha = 128, LoRA Dropout = 0.0, LoRA Trainable Modules = all-linear, Optimiser = AdamW, Learning Rate Scheduler = Linear, Max Grad Norm = 1.0`
Fine-Tuned OpenAI Models (GPT-4o, GPT-4o-mini)	`num train epochs = 3, batch size = 20, learning rate = 1e-5, Optimiser = AdamW, Temperature = 0, Max Tokens = 1024, Top p = 1.0, Frequency Penalty = 0.0, Presence Penalty = 0.0`
GPT-4o ZS and GPT-4o FS	`Max Tokens = 2048, Temperature = 0, Top p = 0.9, Frequency Penalty = 0.0, Presence Penalty = 0.0`
Machine Learning Models	
Estimator	**Hyperparameters**
Logistic Regression (LR)	`penalty=l2, C=1.0, solver=lbfgs, max_iter=100, verbose=0`
Support Vector Machine (SVM)	`C=1.0, gamma=1.0, cache_size=200, max_iter=-1`
Random Forest (RF)	`n_estimators=100, max_depth=10, min_samples_split=2`
AdaBoost	`n_estimators=50, learning_rate=1.0, base_estimator=DecisionTreeClassifier, algorithm=SAMME.R`
CatBoost	`iterations=1000, learning_rate=0.03, depth=6, verbose=True`

Table 13: Parameters for Deep Learning, Transformer-Based, and Machine Learning Models. The first section presents trainable model parameters, while the second section highlights hyperparameters for machine learning models.

Strategy	Prompt	Example
Zero-shot prompting	**Task explanation** **Explicit behavioural guidelines** **Task instance**	**Your task:** Classify a given sentence as either *slang* or *non-slang*. 1. **FIRST line:** ONLY write *slang* or *non-slang* with no extra words or punctuation. 2. **Following lines:** Explain WHY the sentence is classified as slang or non-slang. 3. **Ensure** the response follows this format for accurate extraction. **Task instance:** *Ravens fans, players still salty, got ass kicked last week*. **Expected Answer:** Slang **Reasoning:** The phrase *still salty* is slang for being upset or bitter about a past event. The phrase *got ass kicked* further emphasises informality.
Few-shot prompting	**Task explanation** **Explicit behavioural guidelines** **Example instances** **Task instance**	**Your task:** Classify a given sentence as *slang* or *non-slang*. 1. **FIRST line:** ONLY write *slang* or *non-slang* with no extra words or punctuation. 2. **Following lines:** Explain WHY the sentence is classified as slang or non-slang. 3. **Ensure** the response follows this format for accurate reasoning extraction. **Example instance:** *Man, that chronic had me feeling way too relaxed last night*. **Answer:** Slang **Reasoning:** The term *chronic* is slang for high-quality marijuana. **Example instance:** *Scientists discovered a new type of germ in the petri dish*. **Answer:** Non-slang **Reasoning:** The word *germ* is used literally in a microbiological context. **Task instance:** *Ravens fans, players still salty, got ass kicked last week*. **Expected Answer:** Slang **Reasoning:** The phrase *still salty* expresses resentment, and *got ass kicked* adds to the informal slang tone.

Table 14: Zero-shot and Few-shot prompting strategies for slang detection.