Fine-tuning BERT for Slang Detection and Identification

Sohaila Abdulsattar Mohammed
New York University
sm10688@nyu.edu

Saamia Shafqat
New York University
ss14758@nyu.edu

Abstract

The dynamic and informal nature of English slang poses significant challenges for natural language processing (NLP) systems, as it often lacks standardized definitions, evolves rapidly, and depends heavily on context. This paper addresses these challenges by fine-tuning BERT, a transformer-based language model, for two key tasks: slang detection and slang identification. We adopt the definition of slang as either entirely new terms created for informal use or existing words that have acquired novel, context-specific meanings. The detection task involves binary classification to determine the presence of slang in sentences, while the identification task locates slang phrases at the token level. A diverse dataset combining written and spoken slang was curated, ensuring robust evaluation across multiple domains. Our fine-tuned BERT model achieved an accuracy of 88% and an F1-score of 0.88 for slang detection, outperforming many traditional models in this task. For slang identification, our fine-tuned BERT model achieved an accuracy of 97% and an F1-score of 0.97, demonstrating significant improvements over other methods in this domain. These results highlight the effectiveness of transformer architectures in tackling the context-dependent and dynamic nature of slang, advancing NLP's ability to handle informal language.

1 Introduction

The English language is in a constant state of evolution, reflecting cultural shifts, social dynamics, and technological advancements. Among its many forms, slang occupies a unique position, serving as both a marker of identity and a dynamic linguistic innovation. Slang can be broadly categorized into two types (Pei et al., 2019):

- **New meanings for existing words:** This occurs when familiar words take on novel, often context-dependent meanings. For example, the word "fire" has evolved to mean "impressive" or "excellent" in informal usage.
- **Newly coined terms:** These are entirely new words or expressions created to convey specific ideas or sentiments. They can include abbreviations like "YOLO" (You Only Live Once) or portmanteaus such as "hangry" (hungry + angry).

These linguistic innovations are especially prevalent on social media platforms like Twitter and Reddit, where character limits and informal interactions drive the adoption of creative, non-standard expressions. This linguistic phenomenon is not confined to the internet only, though; it also pervades spoken language, shaping how people communicate every day. Understanding slang is essential for enhancing natural language processing (NLP) tasks, improving accessibility for diverse audiences, and ensuring the inclusivity of language technologies.

Despite its significance, detecting and identifying slang poses unique challenges due to its informal, context-dependent nature. Current NLP models, typically trained on formal language datasets, struggle with slang, which often lacks standard definitions or consistent spelling. This gap hinders applications such as sentiment analysis, machine translation, and content moderation. Furthermore, non-native speakers, older generations, or individuals outside specific subcultures often find it difficult to interpret slang, creating a barrier to understanding and engagement.

To address these challenges, this paper focuses on two core tasks:

1. **Slang Detection:** This is a binary classification task that determines whether a given sentence contains any slang expressions, focusing solely on identifying their presence or absence.

2. **Slang Identification**: This task focuses on pinpointing the exact location of slang words or phrases in a sentence. It involves labeling each token using the BIO scheme (Begin, Inside, Outside) (Ramshaw and Marcus, 1999) to mark which parts of the text constitute slang and which do not.

While our tasks do not involve differentiating between the two categories of slang defined previously, these definitions have guided our choice of datasets to ensure a diverse and representative corpus. We elaborate on these tasks by leveraging the transformative capabilities of the Bidirectional Encoder Representations from Transformers (BERT) machine learning model (Kenton and Toutanova, 2019). Fine-tuning BERT allows for robust contextual understanding, which is critical for handling the variability and creativity inherent in slang. Through this work, we propose a comprehensive solution that not only advances slang detection and identification but also enhances the adaptability of NLP systems to informal language.

2 Related Work

The task of slang detection and identification has been studied across various domains, leveraging approaches ranging from rule-based systems to modern deep learning frameworks. This section categorizes related work into two areas: approaches focused specifically on slang detection and identification and research employing BERT for linguistically similar tasks, such as metaphor detection and sarcasm identification. We also highlight gaps in existing methods and connect prior research to our proposed work.

2.1 Slang Detection and Identification

Early efforts in slang processing emphasized normalization as a critical step for improving the performance of downstream NLP tasks. For instance, Han and Baldwin (2011) developed a system for lexical normalization of social media text, incorporating spell correction and dictionary lookups to replace slang and misspelled words with their standardized equivalents. Their work laid the foundation for considering context during the normalization process, but it relied heavily on static resources, limiting its applicability to emerging slang terms.

Building on these ideas, Adedamola et al. (2015) proposed a rule-based system to normalize internet slang in social media. Their approach combined tokenization, regular expressions, and slang dictionaries to classify tokens as in-vocabulary (IV) or out-of-vocabulary (OOV), subsequently replacing OOV terms with their standard equivalents. While effective for preprocessing tasks, their system struggled with context-sensitive slang, where meaning depends on usage, and required frequent updates to the slang dictionary to remain relevant.

Later works advanced these ideas using machine learning and deep learning techniques. Pei et al. (2019) addressed the foundational task of automatic slang detection and identification by framing it as a sequence labeling problem. Leveraging Bidirectional Long Short-Term Memory (BiLSTM) models combined with Conditional Random Fields (CRFs) and Multilayer Perceptron (MLP), their models achieved an F1-score of 0.80 for sentence-level detection and 0.50 for token-level identification. To bolster performance, Pei et al. (2019) integrated linguistic features like Pointwise Mutual Information (PMI), Part-of-Speech (POS) tagging, and character-level embeddings. Their feature ablation analysis revealed that POS transformations were the most diagnostic for slang, particularly in cases of syntactic shift, where a slang word was twice as likely to change its grammatical category compared to standard usage.

Rothe et al. (2023) expanded the scope of slang processing by integrating different deep learning models. Their work employed various techniques for the task of slang detection, including Gaussian Naïve Bayes, Random Forests, LSTM, and Bi-LSTM; it also focused on token-level classification using BERT, achieving fine-grained detection of slang terms through the BILOU tagging scheme. Evaluated on the Slangvolution Twitter corpus, which included over 225,000 tagged tweets, their models demonstrated strong performance, with BiLSTM achieving a 95.37% accuracy in detecting slang. The inclusion of BERT for token-level classification further underscored the potential of transformer models for handling nuanced linguistic patterns in slang.

2.2 BERT in Linguistically Similar Tasks

Transformer models like BERT have shown remarkable success in tasks that require a nuanced understanding of language. These tasks often share challenges with slang detection, such as handling context-dependent meanings and variability in usage.

Metaphor detection, for instance, presents similar complexities. Choi et al. (2021) proposed MelBERT, a framework that leverages BERT with linguistic theories such as the Metaphor Identification Procedure (MIP) and Selectional Preference Violation (SPV). MelBERT uses a late-interaction architecture to independently encode a word and its context, enabling the model to capture the semantic gap between literal and metaphorical meanings. Their work demonstrated the potential of combining linguistic theories with BERT to address semantic ambiguity, achieving state-of-the-art results on benchmark datasets like VUA-18 and VUA-20 with F1 scores of 78.5% and 72.3%, respectively. This methodology offers valuable insights for slang detection, where contextual understanding is equally crucial.

Sarcasm detection, another task reliant on contextual cues, also benefits from transformer models. Parameswaran et al. (2021) explored sarcasm target detection using two BERT-based models: TD-BERT and BERT-AEN. By formulating the problem as sequence labeling, they fine-tuned BERT on domain-specific datasets, from sources like Reddit and Twitter, capturing the informal and nuanced language typical of these platforms. Their method achieved a 15.2% improvement on the Reddit data and a 4% improvement on Tweets, in comparison to other recent state-of-the-art systems. Their results also highlighted the importance of pre-training on domain-specific data, a strategy that can be effectively adapted for slang detection, especially when processing emerging or spoken slang.

2.3 Connections to Our Work

Our research builds on these advancements by leveraging the power of transformer-based models, specifically BERT, to tackle the dual tasks of slang detection and identification. Our work integrates a highly diversified dataset that combines both written and spoken language; this multimodal approach addresses a critical gap in existing literature, enabling our model to adapt effectively across contexts and domains where slang usage varies widely.

Previous studies like Pei et al.'s (2019), which utilized the Wall Street Journal Corpus and The Online Slang Dictionary, and Rothe et al.'s (2023), which relied on Twitter datasets, provide valuable insights into the complexities of slang. Building on these works, our approach adopts a multimodal perspective, incorporating both spoken and written slang to better reflect the diverse ways slang is used in real-world communication.

Additionally, our work adapts and fine-tunes BERT to handle the contextual and dynamic nature of slang, drawing inspiration from its successful applications in metaphor detection (e.g., MelBERT) and sarcasm identification (e.g., TD-BERT). These tasks share linguistic complexities with slang detection, such as context-dependence and semantic shifts, which we address through BERT's transformer architecture. By doing so, we demonstrate the adaptability of BERT for identifying slang across diverse modalities.

Further details on the construction and utility of our multimodal dataset, as well as the methods underpinning our approach, will be elaborated in subsequent sections.

3 Methodology

3.1 Leveraging BERT for Slang Detection and Identification

BERT, first introduced by Kenton and Toutanova (2019), has become a cornerstone of modern natural language processing due to its capacity for deep contextual understanding. Unlike traditional word embedding models such as Word2Vec (Mikolov, 2013) or GloVe (Pennington et al., 2014), which generate static embeddings, BERT leverages a transformer-based architecture to produce dynamic embeddings that capture the context of words based on their position in a sentence.

BERT's architecture comprises a multi-layer bidirectional transformer, allowing it to attend to both preceding and following words simultaneously. This bidirectionality is a significant advancement over models like GPT (Radford, 2018), which process text unidirectionally. For tasks like slang detection and identification, where the interpretation of a term depends on surrounding words, this ability to capture nuanced context is indispensable. Additionally, the pre-training objectives of BERT—Masked Language Modeling (MLM) and Next Sentence Prediction (NSP)—equip it with a robust understanding of linguistic structures and relationships, further enhancing its adaptability to informal and ambiguous language.

By leveraging BERT's strengths, we aim to develop a system that identifies slang with precision and adapts seamlessly to the evolving and multimodal characteristics of informal language. This approach positions BERT as a key enabler in ad-

vancing the processing and understanding of informal communication in NLP.

3.2 Datasets

3.2.1 Slang Detection Dataset

The dataset for the slang detection task consists of binary-labeled sentences, where each sentence is annotated with a 1 if it contains slang and a 0 otherwise. This dataset was constructed to ensure diversity in sources and a balance between positive and negative examples. Our primary sources include the OpenSub-Slang dataset by Sun et al. (2024), an open-source Gen Z Slang dataset from Hugging Face (MLBtrio, 2024), and the Wall Street Journal (WSJ) Penn Treebank dataset (Marcus et al., 1999), each contributing unique characteristics to the corpus.

The OpenSub-Slang dataset, derived from the Open Subtitles Corpus, provided 25,000 human-annotated sentences. Among these, 7,488 sentences were labeled as containing slang, while 17,512 were labeled as non-slang. Each sentence in this dataset is associated with metadata, such as the region and year of production of the movie from which it originates. Slang-containing sentences were further annotated with a confidence score (out of 3) to reflect agreement among annotators. For high-confidence cases, additional annotations included detailed definitions and one-word paraphrases of the slang terms. To ensure reliability, we filtered out sentences with confidence ratings below 2, resulting in a subset of 2,256 positively labeled sentences. Notably, this dataset primarily contributed examples of slang where existing words have acquired new meanings, reflecting the linguistic evolution observed in spoken contexts.

The Hugging Face Gen Z Slang dataset augmented our positive instances with 1,779 sentences sourced from a combination of a Kaggle dataset of social media slang and acronyms, as well as a curated Gen Z slang collection compiled by scraping public websites, Google searches, and TikTok content. This dataset primarily provided examples of newly coined slang terms, including acronyms and creative expressions that have emerged in informal digital communication.

To balance the dataset with negative (non-slang) examples, we incorporated the Open Subtitles non-slang dataset, supplemented by sentences from the WSJ Penn Treebank corpus. Approximately 50% of the negatively labeled sentences were sourced from WSJ to enhance diversity. We further curated the dataset to ensure that the sentence length distribution for non-slang examples matched that of the slang-containing sentences, reducing potential biases based on sentence length when training our model. The final dataset was randomly shuffled to intermix binary labels and again to prevent any sequential patterns that could bias the model during training.

Overall, our curated dataset for slang detection includes 8,070 sentences, half of which are positively labeled and the other half negatively labeled. By combining datasets from diverse domains and ensuring high-quality annotations, we created a comprehensive resource capable of capturing the linguistic nuances of slang across written and spoken modalities.

3.2.2 Slang Identification Dataset

For the slang identification task, the positive instances from the slang detection dataset were further processed to annotate the exact location of slang terms within sentences using the BIO tagging scheme. This process allowed for token-level granularity in identifying slang expressions to enable our model to pinpoint slang phrases effectively.

Each sentence was first tokenized into individual words, with additional splitting on punctuation to handle cases where slang terms span multiple words or are embedded within larger phrases. Tokens identified as part of a slang phrase were tagged as follows:

- **B (Beginning)**: The first token of a slang expression.
- **I (Inside)**: Subsequent tokens within the same slang expression.
- **O (Outside)**: Tokens not part of any slang expression.

Approximately 5.5% of the sentences (228 instances) required manual intervention due to complexities in slang usage or tokenization issues. These included cases where slang phrases were ambiguous or not easily segmented by automated tools. Manual annotation ensured consistency and accuracy across the dataset, addressing edge cases effectively.

The resulting BIO-tagged dataset consists of 4,035 sentences with detailed annotations for token-level classification, forming the foundation for training the slang identification model.

3.3 Model Fine-Tuning

To ensure robust training and evaluation, the dataset for each of the tasks was divided into two distinct parts: 85% of the data was allocated for training and cross-validation, while the remaining 15% was set aside as a separate test set. The test set was not used during the training or validation process, serving solely to evaluate the final model's generalization capability on unseen data.

Within the 85% training portion, we employed K-fold cross-validation with five folds to validate model performance. For each fold, one subset of the data was designated as the validation set, while the remaining subsets served as the training set. By rotating through all five folds, every data point contributed to both training and validation. This method ensured that the model was rigorously evaluated across different subsets of the training data.

Cross-validation was particularly important for these tasks because it allowed us to maximize the utility of the dataset, which is relatively small compared to datasets used for other large-scale NLP tasks. By iteratively training and validating across multiple folds, we reduced the risk of overfitting and ensured that the final model was robust and reliable. The separation between the cross-validation set and the held-out test set further reinforced the evaluation pipeline, ensuring that performance metrics on the test set truly reflected the model's ability to generalize to unseen data. This combination of cross-validation and a dedicated test set provided a comprehensive framework for assessing the model's stability and adaptability.

3.3.1 Detection Task

For the slang detection task, we fine-tuned the BERT large uncased model as a binary classification model to determine whether a given sentence contains slang. To optimize computational efficiency and minimize the risk of overfitting, the base layers of the pre-trained BERT model were frozen during training. This approach preserved the general linguistic patterns learned during BERT's extensive pre-training on large corpora, while allowing the model to focus on adapting to the specific slang detection task. In contrast, the pooler layers, responsible for generating task-specific embeddings, were unfrozen and updated during training to fine-tune the model to the slang detection objective.

A fully connected layer was added as a classification head atop the BERT architecture, producing binary predictions corresponding to the two labels: slang-containing (1) and non-slang (0). This classification head allowed the model to map the contextualized embeddings produced by BERT to the binary output space.

We utilized the AutoTokenizer from Hugging Face and initialized it with the BERT large uncased model to preprocess text into tokenized input suitable for BERT. Text sequences were tokenized using the tokenizer, applying truncation and padding to ensure uniform input length for the model. Tokenized datasets were formatted as PyTorch tensors, compatible with the Trainer API provided by Hugging Face. The model was then finally trained for four epochs for each fold as part of the K-fold cross-validation strategy previously explained. The flow of the BERT fine-tuning process for the detection task is summarized in Figure 1.

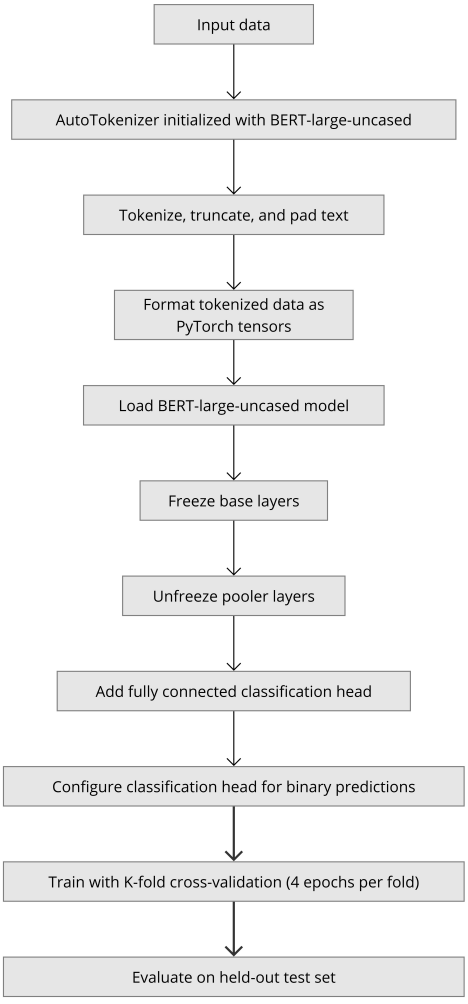

Figure 1: Flow of BERT Finetuning for Detection Task

3.3.2 Identification Task

Fine-tuning BERT for the slang identification task involved adapting the BERT large uncased model for token classification, ensuring precise labeling of slang tokens within sentences according to the BIO tagging scheme. This task involved addressing the complexities of BERT's WordPiece tokenization, which is a key challenge in named entity recognition (NER) and similar sequence labeling tasks. Unlike traditional word tokenization, BERT's WordPiece tokenization splits words into subwords. To ensure accuracy, we decided to propagate the original BIO label of a word to all its subwords, such that each token retains the correct label. We utilized a specialized tokenization function to handle this process, ensuring that word-level labels were mapped consistently to subword tokens. While slightly slower in processing time, this method helps model performance by maintaining alignment between labels and input tokens.

After tokenizing each sentence using BERT's WordPiece tokenizer, special tokens ([CLS] and [SEP]) were added to indicate the start and end of the sequence, as required by BERT. Sentences were then padded or truncated to fit the maximum input length supported by the model. An attention mask was created alongside the tokenized inputs to differentiate between meaningful tokens and padding. The BIO labels were carefully aligned with the tokenized outputs using the label propagation technique described previously, ensuring that subwords inherited the tags of their parent words. Finally, the processed tokenized inputs, attention masks, and aligned labels were converted into PyTorch tensors to prepare them for training.

For the model architecture, we employed the "BertForTokenClassification" implementation from the Hugging Face library, initializing it with pre-trained weights from the BERT large uncased model. This model is specifically designed for token-level tasks, with a classification head added on top of the BERT base layers to predict labels for each token. The classification head was initialized with randomly generated weights, which were fine-tuned alongside the pre-trained weights using the labeled dataset. This design allowed BERT to leverage its pre-trained linguistic knowledge while adapting to the specific requirements of the slang identification task.

The training process was carried out using PyTorch, and we employed K-fold cross-validation as previously explained. The model was trained for six epochs per fold, and later finally evaluated on the held-out test set. Figure 2 summarizes the flow for fine-tuning BERT on the identification task.

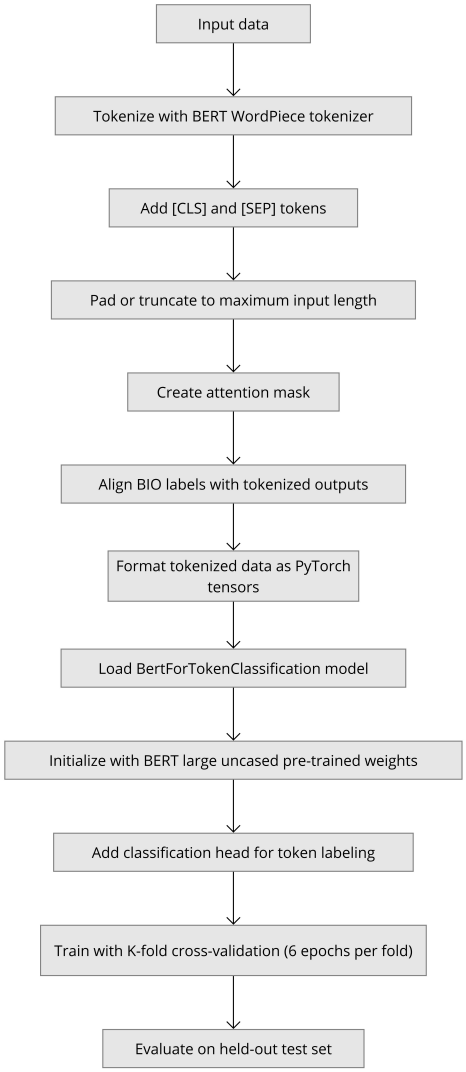

Figure 2: Flow of BERT Finetuning for Identification Task

4 Evaluation

For our models, we employed four standard evaluation metrics—accuracy, precision, recall, and F1-score—each providing unique insights into the model's performance. For the binary classification (slang detection) task, these metrics were calculated at the sentence level. For the BIO-tagging (slang identification) model, evaluation was performed at the token level, allowing for fine-grained analysis of the model's ability to identify specific slang phrases.

4.1 Accuracy

Accuracy is the ratio of correctly predicted instances to the total number of predictions. It provides a general overview of the model's correctness:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

While accuracy is a helpful summary metric, it can be misleading when the dataset is imbalanced. For example, often in BIO tagging, majority of tokens are labeled "O". A model biased toward predicting "O" could achieve high accuracy without effectively detecting "B" or "I" tokens. Hence, additional metrics like precision, recall, and F1-score are crucial for a comprehensive evaluation.

4.2 Precision

Precision measures the proportion of correctly predicted positive instances out of all instances predicted as positive:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

High precision would indicate fewer false positives.

4.3 Recall

Recall is the proportion of correctly predicted positive instances out of all actual positive instances:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

High recall means the model successfully captures most of the relevant items.

4.4 F1-Score

The F1-score is the harmonic mean of precision and recall, providing a balanced measure that accounts for both false positives and false negatives:

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

5 Results

The performances of our BERT fine-tuned slang detection and identification models were evaluated on the test sets, showcasing their effectiveness in handling the complexities of slang across diverse contexts. The results are detailed below, with comparisons to previous studies highlighting the strengths and limitations of our approach. Unless otherwise specified, all scores mentioned in this section refer to weighted averages, which provide a balanced measure that accounts for class imbalances.

5.1 Slang Detection Model

Our BERT-based slang detection model achieved an overall accuracy of 88% on the final test set, with an F1-score of 0.88. These results reflect our model's robust ability to distinguish between slang-containing and non-slang sentences, leveraging the contextual understanding provided by BERT and the diversity of our dataset. Detailed performance metrics for the two classes—"Not Slang" (0) and "Slang" (1)—are summarized in Table 1. Table 2 compares these metrics to a simple baseline system that tags all sentences as non-slang (0). Additional insights into the model's results are detailed in the confusion matrix in Figure 3.

Class	**Precision**	**Recall**	**F1-Score**
Not Slang (0)	0.89	0.86	0.87
Slang (1)	0.87	0.90	0.88
Macro Avg	0.88	0.88	0.88
Weighted Avg	0.88	0.88	0.88
Accuracy		0.88	

Table 1: Fine-tuned BERT Model Performance Metrics for Slang Detection

Class	**Model**			**Baseline**		
	Precision	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
Not Slang (0)	0.89	0.86	0.87	0.49	1.00	0.66
Slang (1)	0.87	0.90	0.88	0.00	0.00	0.00
Macro Avg	0.88	0.88	0.88	0.25	0.50	0.33
Weighted Avg	0.88	0.88	0.88	0.24	0.49	0.33
Accuracy		0.88			0.49	

Table 2: Comparison of BERT Model vs All Non-Slang Baseline for Slang Detection

When compared to results reported by Rothe et al. (2023), our model demonstrated competitive performance. Specifically, it outperformed the Gaussian Naïve Bayes approach, which achieved an F1-score of 0.815 and an accuracy of 78.33%. However, our model's performance did not surpass the Random Forest Classifier (F1-score of 0.9695, accuracy of 96.01%) and LSTM-based models (F1-scores from 0.9610 to 0.9639, accuracy scores from 95.00% to 95.37%) also reported in Rothe et al.'s (2023) study. This indicates that while our approach is highly effective, other methods may

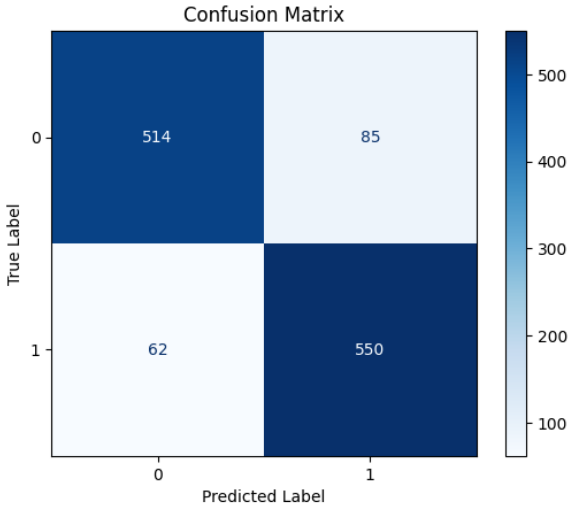

Figure 3: Confusion Matrix for BERT Slang Detection Model

outperform BERT in specific settings, particularly when trained on larger, domain-specific datasets.

Comparing our results with Pei et al.'s (2019), our model demonstrated superior performance against various BiLSTM-based configurations. Pei et al.'s (2019) BiLSTM-MLP models achieved F1-scores ranging from 0.6469 to 0.7931, depending on feature selection, while BiLSTM-CRF models achieved F1-scores between 0.7440 and 0.7971. Even the best-performing BiLSTM-CRF configuration, with full feature inclusion, achieved an F1-score of 0.7971, below our model's 0.88. These comparisons highlight the robustness of our transformer-based approach for detecting slang across diverse domains.

Table 3 provides a summary of our model's performance compared to the previously reported results by Rothe et al. (2023) and Pei et al. (2019), presenting the F1-scores for each.

5.2 Slang Identification Model

Our BERT-based BIO-tagging model for slang identification demonstrated exceptional performance, achieving an overall accuracy of 97%. Its weighted average F1-score across token-level labels was 0.97, while the macro average F1-score was 0.82. These strong results, detailed in Table 4, indicate high precision and recall for "B" and "O" tokens, though performance on "I" tokens was lower, suggesting room for improvement in handling multi-word slang expressions. Table 5 compares our BERT model's results to a simple baseline system that tags all tokens as "O" (Outside).

Model	F1 Score
Gaussian Naïve Bayes	**0.8150**
Random Forest Classifier	**0.9695**
LSTM	**0.9639**
Bidirectional LSTM	**0.9610**
BiLSTM-MLP (POS+POSp)	0.6469
BiLSTM-MLP (POS+POSp+POSt+PMI)	0.7162
BiLSTM-MLP (POS+POSp+PMI boost)	0.7475
BiLSTM-MLP (full features)	0.7931
BiLSTM-CRF (POS+POSp)	0.7440
BiLSTM-CRF (POS+POSp+POSt+PMI)	0.7787
BiLSTM-CRF (POS+POSp+POSt+PMI boost)	0.7797
BiLSTM-CRF (full features)	**0.7971**
Our Fine-tuned BERT model	**0.8820**

Table 3: Comparison of Slang Detection F1-Scores for Various Models

Additional details on the model's results are also provided in the confusion matrix in figure 4.

Class	Precision	Recall	F1-Score
B (Beginning)	0.89	0.95	0.92
I (Inside)	0.65	0.46	0.54
O (Outside)	0.99	0.98	0.99
Macro Avg	0.85	0.80	0.82
Weighted Avg	0.97	0.97	0.97
Accuracy		0.97	

Table 4: Metrics for Slang Identification

Class	Model			Baseline		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
B (Beginning)	0.89	0.95	0.92	0.00	0.00	0.00
I (Inside)	0.65	0.46	0.54	0.00	0.00	0.00
O (Outside)	0.99	0.98	0.99	0.86	1.00	0.92
Macro Avg	0.85	0.80	0.82	0.29	0.33	0.31
Weighted Avg	0.97	0.97	0.97	0.73	0.86	0.79
Accuracy		0.97			0.86	

Table 5: Comparison of BERT Model vs All Non-Slang Baseline System for Slang Identification

When compared to Pei et al.'s (2019) results on the slang identification task, our model showed substantial improvements. Pei et al.'s (2019) highest-performing BiLSTM-MLP model with full features achieved an F1-score of 0.4985, lower than our F1-score of 0.97.

Pei et al.'s (2019) models also struggled with recall, achieving values between 0.3172 and 0.4612 across configurations. In contrast, our model achieved much higher recall, with a macro average of 0.80 and a weighted average of 0.97. For preci-

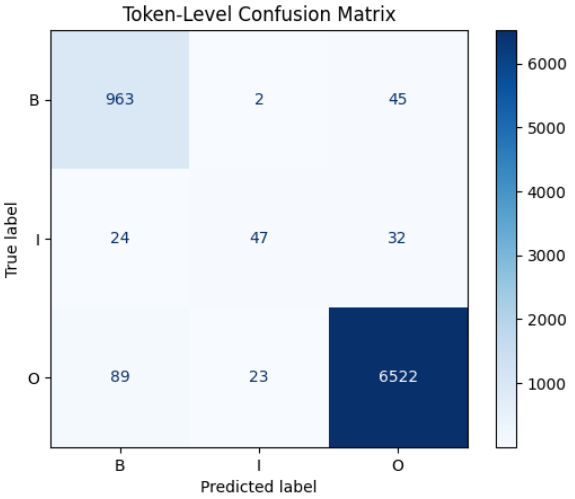

Figure 4: Confusion Matrix for BERT Slang Identification Model

sion, Pei et al.'s (2019) BiLSTM-MLP model with POS features achieved the highest value 0.6240, while our model again outperformed it with macro and weighted averages of 0.85 and 0.97, respectively. These results demonstrate the effectiveness of our BERT-based approach in reliably identifying slang phrases.

Table 6 summarizes the comparison of our identification model's performance against Pei et al.'s (2019), highlighting precision, recall, and F1-scores.

Model (features)	Precision	Recall	F1-score
BiLSTM-MLP (POS+POSp)	0.6240	0.3172	0.4206
BiLSTM-MLP (POS+POSp+POSt+PMI)	0.6172	0.3864	0.4753
BiLSTM-MLP (POS+POSp+POSt+PMI boost)	0.5967	0.3975	0.4771
BiLSTM-MLP (full features)	0.5423	0.4612	0.4985
BiLSTM-CRF (POS+POSp)	0.5666	0.3712	0.4485
BiLSTM-CRF (POS+POSp+POSt+PMI)	0.5763	0.4183	0.4847
BiLSTM-CRF (POS+POSp+POSt+PMI boost)	0.5954	0.4280	0.4980
BiLSTM-CRF (full features)	0.5499	0.4501	0.4950
Our Fine-tuned BERT model	0.9700	0.9700	0.9700

Table 6: Comparison of Slang Identification Precision, Recall, and F1-Scores for Various Models

6 Conclusion

This paper presents an effective approach for detecting and identifying English slang by fine-tuning BERT, a transformer-based architecture well-suited for context-sensitive tasks. Our models demonstrated strong performance, with the slang detection model achieving an accuracy of 88% and an F1-score of 0.88, and the BIO-tagging model for slang identification achieving an accuracy of 97% and an F1-score of 0.97. These results highlight our models' robustness in distinguishing slang-containing sentences and reliably identifying slang phrases at the token level.

When compared to existing models, our approach showed clear strengths. Against BiLSTM-based methods reported by Pei et al. (2019), our models consistently outperformed in both the detection and identification tasks. Similarly, our detection model exceeded the performance of Gaussian Naïve Bayes reported by Rothe et al. (2023). However, our detection model fell short when compared to the highest-performing LSTM-based and Random Forest models in Rothe et al.'s (2023) study, which benefited from training on more extensive domain-specific data. These comparisons underscore the competitive strengths of our transformer-based approach and point to the possibility for further exploration of hybrid methodologies that can leverage the strengths of multiple models.

Despite the successes of our models, some limitations were evident. The detection model occasionally struggled with ambiguous or highly context-dependent slang, particularly in complex sentences. The BIO-tagging model, while performing strongly overall, showed weaker performance on "Inside" (I) tokens of multi-word slang phrases, indicating challenges in handling internal components of slang expressions. These challenges highlight the need for further improvements to address subtle linguistic variations and enhance the model's sensitivity to multi-word slang expressions.

A key limitation of our study is the lack of larger open-source datasets, which also constrained our ability to leverage the full capabilities of a pipeline model. As a result, detection and identification were treated as separate tasks using the same dataset, rather than integrating outputs from detection into identification. Future work could focus on building a comprehensive pipeline that connects the two models, allowing for end-to-end slang processing. Incorporating a normalization step to convert slang-containing sentences into their formal equivalents would also further enhance the adaptability of NLP systems in real-world applications.

In conclusion, our work highlights the potential of transformer-based architectures like BERT for processing slang, advancing NLP's ability to handle the informal and evolving nature of English language. By addressing key challenges and leveraging BERT's contextual understanding, this study sets a strong foundation for future research in developing modular, scalable, and comprehensive sys-

tems for slang detection and identification.

References

Adedoja A Adedamola, Abiodun Modupe, and Olumuyiwa J Dehinbo. 2015. Development and evaluation of a system for normalizing internet slangs in social media texts. *proeedings of WCECS*.

Minjin Choi, Sunkyung Lee, Eunseong Choi, Heesoo Park, Junhyuk Lee, Dongwon Lee, and Jongwuk Lee. 2021. Melbert: Metaphor detection via contextualized late interaction using metaphorical identification theories. *arXiv preprint arXiv:2104.13615*.

Bo Han and Timothy Baldwin. 2011. Lexical normalisation of short text messages: Makn sens a# twitter. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 368–378.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2. Minneapolis, Minnesota.

Mitchell P. Marcus, Beatrice Santorini, Mary Ann Marcinkiewicz, and Ann Taylor. 1999. Treebank-3. Web Download. LDC Catalog No.: LDC99T42.

Tomas Mikolov. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 3781.

MLBtrio. 2024. Gen Z slang dataset. https://huggingface.co/datasets/MLBtrio/genz-slang-dataset. [Accessed 14-12-2024].

Pradeesh Parameswaran, Andrew Trotman, Veronica Liesaputra, and David Eyers. 2021. Bert's the word: Sarcasm target detection using bert. In *Proceedings of the 19th Annual Workshop of the Australasian Language Technology Association*, pages 185–191.

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.

Alec Radford. 2018. Improving language understanding by generative pre-training.

Lance A Ramshaw and Mitchell P Marcus. 1999. Text chunking using transformation-based learning. In *Natural language processing using very large corpora*, pages 157–176. Springer.

Mrudula Rothe, Ritika Lath, Dishant Kumar, Parth Yadav, and Amit Aylani. 2023. Slang language detection and identification in text. In *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–5. IEEE.

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*.