Proceedings of the Third International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT-2025)
IEEE Xplore Part Number: CFP25CV1-ART; ISBN: 979-8-3315-2754-9

2025 3rd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT) | 979-8-3315-2754-9/25/$31.00 ©2025 IEEE | DOI: 10.1109/IDCIOT64235.2025.10914742

A Hybrid Approach for Language Variety Prediction using BERT and T5 Embeddings

Karunakar Kavuri
Associate Professor, Department of CSE
Swarnandhra College of Engineering and Technology,
Narasapuram, AP, India.
karunakar.mtech@gmail.com

T. Raghunadha Reddy
Associate Professor, Department of CSE
Matrusri Engineering College,
Hyderabad, Telangana, India.
drtrnreddy@gmail.com

Archana Gelli
Assistant Professor, Department of CSE
Swarnandhra College of Engineering and Technology,
Narasapuram, AP, India.
ksj.archana@gmail.com

B H D D Priyanka
Assistant Professor, Department of IT
S R K R Engineering College,
Bhimavaram, AP, India.
priyanka@srkrec.ac.in

***Abstract*—A crucial component of author profiling is language variety prediction, which looks for stylistic and regional differences in language usage in written texts. Applications for this activity are substantial in fields including localization, forensics, and customized marketing. Conventional methods depend on manually created features or embeddings from separate models, which might not fully represent the nuanced linguistic differences. In order to improve language variety prediction performance, we presented a hybrid model in this work that integrates statistical features and embeddings from two sophisticated transformer models, such as T5 (Text-to-Text Transfer Transformer) and BERT (Bidirectional Encoder Representations from Transformers). Different statistical features identified in this work are Lexical features, Syntactic features, Orthographic and Phonological features, Stylistic features, Semantic features, Pragmatic and discourse-level features and Temporal features. The embeddings from both models are projected into a shared dimensions space, averaged, appended, and utilized as input to a downstream classifier in order to accomplish effective integration. This method makes use of T5's generative capabilities and BERT's contextual richness to better capture syntactic and semantic patterns in text. The English language diversity dataset from the PAN 2017 competition is used to test the suggested framework. According to experimental evaluations, the hybrid embedding technique achieves better accuracy and F1-score for language variety prediction, outperforming individual transformer models by a wide margin. When compared to other models for language variety prediction, the XGBoost classifier with Statistical Features, BERT, and T5 embeddings performed better in terms of F1-Score (93.4%) and Accuracy (93.5%). These results demonstrate the importance of including complimentary transformer embeddings to improve author profiling tasks' state-of-the-art performance.**

Keywords—Bidirectional Encoder Representations from Transformers, Text-to-Text Transfer Transformer, Language Variety Prediction, Word Embeddings, Author Profiling, PAN Competition 2017.

I. INTRODUCTION

Deducing demographic characteristics from textual data, such as gender, age, native language, or nationality, is known as author profiling [1]. As a component of author profiling, language variety prediction focuses on identifying an author's native or regional language from their written texts. Finding subtle language characteristics, such as dialects, slang, and structural patterns that represent the author's linguistic diversity or cultural background, entails examining linguistic features. It is especially crucial in situations where minute variations in language use can reveal information about the author's cultural or geographic origins. Predicting language variety in this field aids in addressing a number of applications, including sociolinguistics, marketing and personalization, and forensic security. Predicting language varieties is a fundamental challenge in programs such as the PAN competition's Author Profiling shared tasks [2], which encourage creativity in recognizing language variants from user-generated content like tweets. Researchers hope to improve security and bridge cultural gaps, communication applications, and computational linguistics by refining profiling techniques.

In the past, language variety prediction has been accomplished by combining machine learning and Natural Language Processing (NLP) techniques [3], using stylistic and lexical features such as word choice, syntax, and spelling variations to distinguish between language varieties and dialects [4], embedding models such as word embeddings (e.g., BERT, Word2Vec) to capture contextual and semantic nuances of text, and machine learning classifiers like Random Forests (RFs) and Support Vector Machines (SVMs) to analyse the features for language variety prediction [5].

To distinguish an author's nativity language from their written texts, certain stylistic, linguistic, and contextual patterns reflecting the author's regional or native language characteristics are analysed, and NLP and machine learning (ML) techniques are used to interpret and classify these patterns. Language variety prediction datasets have shown a number of significant differences between texts from various regional varieties, which can be categorised into syntactic, grammatical, lexical, and cultural aspects, offering useful features for correctly predicting an author's language variety.

Lexical Variations: Dissimilar regions of the same language frequently use different vocabulary. For instance, American English uses "cookie," while British English uses

"biscuit." Similarly, Australian English frequently uses unusual terms like "arvo" for "afternoon," which is rarely used in British or American contexts. These lexical choices provide clear linguistic markers.

Regional variations in spelling offer distinct linguistic signals in Spelling Differences. Words like "colour" and "favour," for example, are spelt with an extra "u" in British English, while "u" is dropped in American English to become "colour" and "favour." Additionally, American English punctuation is placed inside quotation marks, whereas British texts support punctuation outside of them.

Regional differences exist in syntactical and grammatical choices. In lines like "I have just eaten," for instance, British authors typically use the present perfect tense, whereas American English writers are more likely to use "I just ate." For predicting linguistic diversity, this variation in verb tense usage is a trustworthy indicator and a recognisable geographical characteristic.

Regional idioms or culturally distinctive references, such local slang in Irish literature or "eh" in Canadian English, are frequently used by Irish or Canadian writers in the context of regional texts. These environmental cues can effectively indicate the author's place of origin.

In Informality and Online Speech, the evolution of digital communication has contributed to more pronounced regional idiosyncrasies in informal writing. Slang, like "innit" in British conversations versus "y'all" in American texts, is a good example of regional online speech patterns. Furthermore, variations in the casual expressions and the use of emojis contribute to distinguishing among national dialects on social media platforms.

This paper is organized in 5 sections. Section 2 discuss about different research works proposed for language variety prediction. The proposed methodology explained in section 3 with the description of dataset characteristics and evaluation measures. The experimental results of proposed approach are presented and discussed in section 4. The section 5 specifies the conclusions of this work with future enhancements to this work.

II. Literature Review

Language variety prediction is an important aspect of author profiling, which has gained increasing attention in NLP due to its applications in marketing, security, and sociolinguistics. Over the years, researchers have explored a diversity of methods to address the challenge of identifying the dialect or language variety of an author based on their textual data. Early approaches mainly relied on statistical patterns and lexical features, including character-level frequency distributions and n-gram models. These techniques were later augmented with syntactic and semantic analyses to capture deeper linguistic variations among language varieties. With advancements in machine learning, predictive models such as Support Vector Machines (SVM) and Random Forests have shown promise in differentiating closely related dialects and languages. More recently, deep learning approaches, including word embeddings like contextualized models like BERT and Word2Vec, have significantly improved performance by capturing both global and local text dependencies. Shared tasks such as those organized by the DSL and PAN Challenge have further stimulated innovation by providing benchmarks and annotated datasets for language variety prediction.

Methods for predicting an author's language variety based on their English writing style were investigated by Raghunadha Reddy, T. et al., [6]. In order to enable classification algorithms to forecast language diversity, they devised a Profile-specific Document Weighted (PDW) technique that uses document weights and term weights to express text features. To enhance the representation of linguistic traits specific to each language variety, the PDW model uses supervised measurements to determine document and term weights. In order to create document representations, the authors used the PAN17 Twitter corpus, which included 360,000 postings that were evenly split among six native English dialects. They then extracted the 8,000 most frequently used words. The effectiveness of a number of classifiers, such as Random Forest, Logistic Regression, and Naïve Bayes, in predicting linguistic varieties was assessed. When tested using the entire feature set of 8,000 words, the Random Forest classifier outperformed other methods and current approaches, achieving the greatest accuracy of 88.57%.

Sameeah Noreen Hameed et al., investigated [7] language variety identification, which is a task within NLP aimed at distinguishing regional variations of the same language based on grammatical, lexical, and semantic differences. The authors focused on improving the state-of-the-art methods by incorporating transfer learning techniques (ULMFiT and BERT) and comparing them against conventional deep learning models (GRU, Bi-LSTM, CNN, and ensemble methods). Transfer learning approaches, especially ULMFiT, consistently outperformed deep learning models for both multi-class and binary tasks. Among deep learning methods, GRU and CNN showed better performance than Bi-LSTM. The Portuguese language variety task attained the highest accuracy of 98.03% with ULMFiT.

Daniel Escobar-Grisales et al., concentrated [8] on recognizing demographic traits like language variety (LV) and gender in texts written in Informal language data (tweets) and formal language data (call-center conversations). Authors employed Recurrent Neural Networks (RNNs), particularly Bi-LSTM networks, and CNNs to categorize demographic traits. They used PAN17 corpus (tweets from Spanish speaking countries) as informal texts, and manually transcribed call-center conversations from Colombia, focusing on regional dialects (e.g., Bogotano vs. Antioqueño) as formal texts. The proposed model attained 92% accuracy for language variety recognition in informal scenarios, and 72% accuracy attained for dialect recognition in formal scenarios.

Swapna M et al., concentrated [9] on improving author profiling techniques by introducing Contextually Propagated Term Weights (CPTW). They aimed to detect author attributes like age, gender, and language variety using an improved term-weighting method which incorporates contextual semantics into feature representation. The CPTW method assigns weights to terms based on their contextual neighbourhood in an embedding space. They utilized PAN 2015 to 2018 datasets for tasks like age and gender prediction across multiple languages. According to the experimental results, gender prediction attained the highest accuracy of 93.76% with CPTW-ICF using SVM, XGBoost excelled in

age prediction with 87.45% accuracy and Random Forest performed best for language variety prediction with an accuracy of 85.27%.

Marcos Zampieri et al., addressed [10] the challenge of distinguishing among closely related language varieties utilizing an evaluation framework and novel dataset. They introduced DSL True Labels (DSL-TL), which is a human-annotated dataset for classifying language varieties. Authors developed DSL-TL with 12,900 human-annotated instances across three languages and evaluated multiple models, like traditional (Naive Bayes variations, including Adaptive Naive Bayes), advanced transformer-based models, and Deep learning models like XLM-R, mBERT, and XLM-R-LD. Authors demonstrated that Naive Bayes-based systems often performed comparably better than deep learning models for this task.

Bassem Bsir et al., explored [11] the application of fine-tuned transformer models for predicting demographic attributes like gender from Arabic social media content. The goal of their task is investigating how fine-tuning a transformer-based model named as Ara-BERTv2-large can improve accuracy in author profiling (AP) tasks, especially for gender prediction. Proposed Ara-BERTv2-large method attained 79.7% accuracy for gender prediction, outperforming general multilingual model of XLM-RoBERTa (76.8%).

Dominique Brunato et al., introduced [12] a text analysis tool called Profiling–UD. This tool leverages linguistic profiling to analyse linguistic complexity and language variation across multiple languages. The tool extracts features spanning several categories such as morphosyntax, lexical variety, subordination, and syntactic structure. These features provide insights into text characteristics like lexical diversity, sentence length, and dependency structure. Authors discussed numerous case studies showcasing the effectiveness of Profiling–UD in tasks like gender prediction, genre classification, and cross-linguistic studies.

Tommi Jauhiainen et al., provided [13] a comprehensive overview of the language identification (LI) field in digital text. It outlines the techniques, history, applications, and challenges of automatic LI, aiming to consolidate fragmented research across disciplines like machine learning, NLP, and information retrieval. Language Identification (LI) involves determining the language of a text segment or document. Authors focused on written text, excluding spoken LI or non-textual forms like images of written text. Various features for LI are surveyed, including word-level features (frequency, dictionaries, position, and morphological patterns), character-level features (bigrams, n-grams, and gapped bigrams), and chunking and higher-level structures (syllables, morphemes, or phoneme-like units), algorithms include statistical, rule-based, machine learning (e.g., neural networks, SVMs), and hybrid methods. They discussed methods to evaluate LI systems using recall, accuracy, and application-specific metrics.

Chennam Chandrika Surya et al., exploreds [14] the task of identifying linguistic varieties (e.g., regional varieties or dialects) of a language using advanced computational approaches. Authors presented a solution to the 2017 PAN competition's challenge of predicting an author's language variety based on tweets. The proposed method used word embedding models of Word2Vec and BERT. The experiment performed with different machine learning algorithms such as SVM and RF to classify language varieties using document vectors as input. The BERT embedding combined with RF classifier attained the best accuracy of 91.87% for language variety prediction when compared with Word2Vec with RF scored 90.51%, and BERT with SVM scored 90.76%.

Francisco Rangel et al., emphasized [15] an in-depth examination of language varieties while incorporating demographic factors of the authors. It introduces a Low-Dimensionality Statistical Embedding approach to represent textual data and evaluates its effectiveness against the top-performing systems in the Author Profiling task of PAN 2017. The proposed method attained an average accuracy of 92.08%, surpassing the leading team's result of 91.84%. The research also investigates the influence of gender, age, and dataset size on identifying language varieties of Arabic.

III. Ease of Use

The methodology proposed is hybrid approach displayed in Fig. 1. In the proposed hybrid approach, first we collected the Language Variety dataset from PAN 2017 competition. Then, apply suitable pre-processing techniques to remove irrelevant data for analysis. Extract all words from the cleaned dataset and a set of statistical features to differentiate the writing styles of different native language authors. All the words are passed to the BERT and T5 models for generating word embeddings. Merge these two varieties of embeddings for all words. Represent each document with the combined word embeddings of all words that are present in that document. The document finally represented with merging of statistical features and embeddings based document representation. These document vectors are used to train the machine learning algorithm of XGBoost algorithm [16]. The trained model predicts the performance of hybrid approach for language variety prediction.

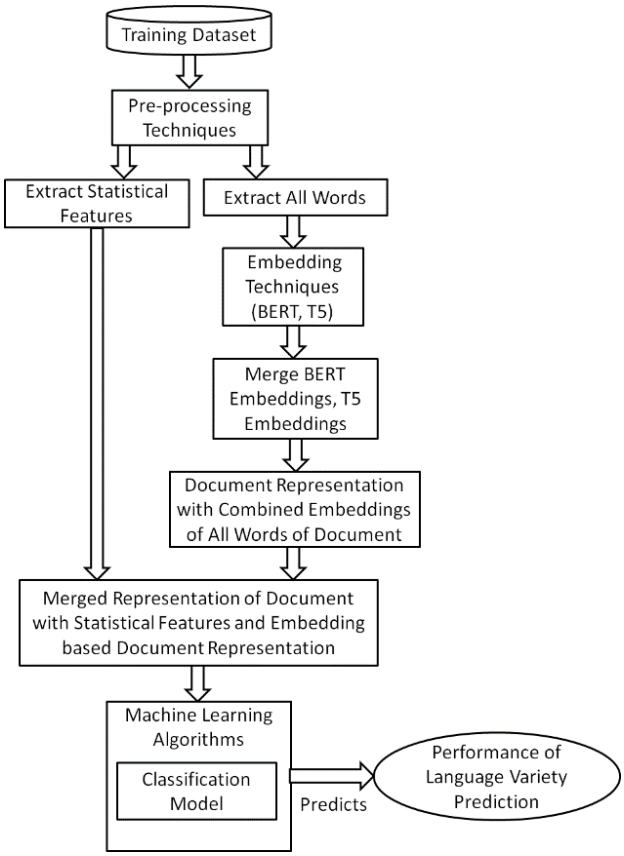

Fig. 1. The Hybrid Approach for Language Variety Prediction

A. Dataset Characteristics

The PAN17 Twitter corpus [2], released in 2017, was utilized for the experimentation in this work. This corpus comprises Twitter posts annotated with variations of English native to particular regions such as Great Britain (GB), Australia (AS), Ireland (IL), Canada (CN), New Zealand (NZ), and the United States (US). The dataset is balanced across these classes, ensuring an equal distribution of samples for comparative analysis. In total, the PAN17 Twitter corpus comprises 360,000 posts, evenly distributed with 60,000 posts per each class. Each regional variant is well-represented, providing a rich linguistic dataset for investigating language variation and its interaction with gender. The labels allow for granular analysis, supporting tasks such as author attribution, linguistic profiling, and dialect recognition.

The corpus serves as a valuable resource for NLP tasks, including sociolinguistic studies and text classification. Its balanced structure facilitates unbiased training and evaluation of machine learning models. Table 1 provides a detailed breakdown of the corpus characteristics, including the number of authors and number of posts per class.

TABLE I. THE CHARACTERISTICS OF PAN 17 TWITTER ENGLISH CORPUS FOR LANGUAGE VARIETY PREDICTION

Number of Authors	Name	Labels	Number of posts	Label distribution	
3600	PAN17 -twitter	Native Language	360000	Ireland	60000
				Canada	60000
				Great Britain	60000
				New Zealand	60000
				United States	60000
				Australia	60000

This corpus is particularly significant for its diverse representation of English dialects, making it appropriate for cross-regional and cross-cultural linguistic analyses. Moreover, the explicit gender annotations enable studies exploring the intersection of sociocultural identity and language use.

The limitations of the dataset are Tweets may lack complete sentences or follow grammatical conventions, dialects like Canadian and American English may have significant overlaps in linguistic features, and small sample size for Fine-Grained Features.

B. Pre-processing Techniquies

Preprocessing is a critical step in preparing the dataset for the language variety prediction task. The different preprocessing techniques that are applied in proposed hybrid model are cleaning the text data by removing URLs, mentions (@username), hashtags, special characters and punctuation. Standardize the text by converting all text to lowercase for consistency. Tokenization of text by using appropriate tokenizers depending on the model (WordPiece tokenizer was used for BERT model, and SentencePiece tokenizer was used for T5 model). Handling Informal Text by replacing informal words or slang with standard equivalents, correction of spellings by using libraries like pyspellchecker. Stop words (e.g., "the", "is", "and") were removed that are not relevant for the task and apply Lemmatization (Converts words to their base or dictionary form) to reduce number of unique words. We extracted all words from the dataset that are available after applying all the preprocessing techniques mentioned.

C. Evaluation Measures

When evaluating the results in author profiling experiments, performance metrics such as precision, accuracy, recall, and F1-score play vital roles in understanding the efficiency of classification algorithms. Among these metrics, F1-score and accuracy are particularly significant for tasks like language variety prediction. Accuracy is the proportion of all predictions that are correct, which is represented mathematically in Equation (1).

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions} \quad (1)$$

Accuracy is a straightforward to interpret and compute, making it a popular choice for assessing overall performance and can reliably represent the performance of classifier when the dataset is balanced (equal distribution of classes).

Language variety prediction often involves multiple classes (e.g., Canadian, Australian, British, etc.). In such cases, high accuracy does not necessarily designate that the classifier performs well across all classes.

The F1-score is the harmonic mean of recall and precision, balancing their trade-off, which is represented in Equation (2).

$$F1-Score=\frac{2\times Precision\times Recall}{\left(Precision+Recall\right)} \quad (2)$$

Precision measures the proportion of correctly predicted instances for a specific class (e.g., how many of the tweets predicted as "British English" are actually British English). Recall (or sensitivity) measures the proportion of true instances of a class that are correctly identified (e.g., how many of the true "British English" tweets were correctly classified). In native language prediction, each regional English variant is treated as a separate class. Computing the F1-score for each class provides a clearer picture of how well the model performs across all categories.

In this work, although accuracy was used to evaluate the performance of the classifiers, incorporating F1-scores could provide deeper insights into the classifier's strengths and weaknesses, particularly for underrepresented dialects.

D. Statistical Features

Advanced statistical features can significantly improve language variety prediction tasks, as they capture nuanced stylistic tendencies and linguistic patterns inherent to different dialects or language variety influences. Below is a list of various statistical features that can be employed for language variety prediction.

Lexical features (Type-Token Ratio (TTR), Word Frequency Distributions, counts of Unigrams, Bigrams, Trigrams, and Skip-Grams, and Word Length Distributions) focus on word-level patterns and distributions, which are often influenced by native language characteristics. Syntactic patterns (Part-of-Speech (POS) Tag Distributions, Function Word Usage, Dependency Parsing Features, Sentence Length and Complexity) are key to identifying regional differences in sentence structure and grammar. Orthographic and Phonological features (Character N-Grams, Spelling Variations, and Punctuation Patterns) influences from native languages or dialects manifest in written forms. Stylistic features (Abbreviation Usage, Capitalization Patterns, Formal vs. Informal Tone, and Emoticon and Emoji Usage) reflect writing habits influenced by cultural or linguistic backgrounds. Semantic features (Sentiment Analysis, Topic Modeling, Word Embeddings, and Named Entity Recognition (NER)) delve into the meanings and associations of words and phrases. Pragmatic and discourse-level features (Discourse Markers, Pronoun Usage, and Cohesion Measures) capture conversational norms and rhetorical strategies. Temporal patterns (Seasonal Linguistic Shifts and Post Timing and Frequency) reflect cultural habits and can be particularly useful for social media-based datasets like Twitter.

E. BERT Embeddings

BERT (Bidirectional Encoder Representations from Transformers) embeddings are highly effective for native language prediction tasks due to their ability to capture contextual and semantic nuances in text [17]. BERT's pre-trained models can be fine-tuned or utilized for feature extraction to improve performance in predicting the language variety or dialect variations [18]. Below is an elaboration on how BERT embeddings can be utilized for native language prediction.

BERT is used for language variety prediction for three purposes such as Contextualized Representations, Pre-trained Knowledge, and Handling Subtle Linguistic Variations. In Contextualized Representations, unlike traditional embeddings (e.g., GloVe and Word2Vec), BERT provides word embeddings that change depending on the surrounding context, making it suitable for capturing regional or cultural nuances. In Pre-trained Knowledge, BERT models are pre-trained on large corpora like BookCorpus and Wikipedia, which can serve as a foundation for learning native language-specific nuances. In Handling Subtle Linguistic Variation, BERT excels at capturing subtle differences in grammar, syntax, and semantics that are critical in language variety or dialect prediction tasks.

To generate embeddings for terms using BERT models, we can leverage either Small Case BERT (BERT-base) or Large Case BERT (BERT-large). Both models use pre-trained weights to produce contextual embeddings, but the primary differences lie in the model's complexity and size. General steps to generate BERT Embeddings are choosing the model variant, pre-processing the input text, embedding generation, and pooling and post-processing. In the first step of choose the Model Variant, select the small Case BERT (BERT-base - It has 12 layers, 12 attention heads, and 768 hidden dimensions per token), or Large Case BERT (BERT-large - It has 24 layers, 16 attention heads, and 1024 hidden dimensions per token, providing richer embeddings at a higher computational cost). In the pre-processing the input text, tokenize the terms using BERT's WordPiece tokenizer, which splits the text into sub-word units, and Convert tokens to lowercase for small case BERT. For large case BERT, maintain the original casing if the task is sensitive to case. In the step of embedding generation, pass the tokenized text through the selected BERT model, extract the embeddings from one or more layers like CLS token embedding (represents the entire input sequence which is useful for sentence or document-level tasks), Word or sub-word embeddings (use the output embeddings for individual tokens), layer aggregation (combine outputs from multiple layers for richer representations using sum, mean, or concatenation). In pooling and post-processing step, apply average-pooling, max-pooling, or other techniques to convert token embeddings into a fixed-length vector representing the document or term. Embedding generation uses transformer-based architectures (BERT and T5) which are represented in Equation (3) and (4).

$$E_B=Transformer_{BERT}\left(T\right) \quad (3)$$

Where, T is tokenized input, and E_B represents contextualized embeddings.

$$E_T=Transformer_{T5}\left(T\right) \quad (4)$$

Where, the encoder generates embeddings optimized for semantic understanding.

Feature Fusion: Concatenation of statistical features S, BERT embeddings E_B, and T5 embeddings E_T, which is represented in Equation (5).

$$F=[S \oplus E_B \oplus E_T] \quad (5)$$

where $\oplus$ denotes feature concatenation.

F. T5 Embeddings

T5 (Text-to-Text Transfer Transformer) embeddings are powerful tools for language variety prediction tasks due to their ability to model both contextual semantics and generate high-quality embeddings through their sequence-to-sequence architecture [19]. Unlike other transformer-based models like BERT, T5 treats all NLP tasks as text-to-text problems, making it versatile for text classification, including language variety prediction.

T5 embeddings are utilized for language variety prediction for three purposes such as pre-trained knowledge, contextual representations, and flexibility. In pre-trained knowledge, T5 is pre-trained on large datasets (e.g., C4 corpus) and captures rich linguistic features that are useful for recognizing subtle language variety influences. In **c**ontextual representations, T5 embeddings are contextualized, meaning sentence or word representations change depending on their linguistic context. In Flexibility, T5's text-to-text framework allows language variety prediction to be framed as a classification or generation task.

Generating embeddings for terms using the T5 model involves leveraging its encoder outputs. Unlike BERT, T5 is a sequence-to-sequence transformer pre-trained on a text-to-text paradigm, which means it transforms input sequences into meaningful embeddings as part of its output generation. The steps to generate T5 embeddings are load pre-trained T5 model, input tokenization, prepare input representations, generate encoder output, aggregate term embeddings, and post-processing. In load pre-trained T5 model step, use a pre-trained T5 model from a library such as Hugging Face (t5-small, t5-base, or t5-large depending on the resources and use case). In input tokenization step, T5 uses a Sentence Piece tokenizer to tokenize the input text (sentences or terms) into sub-words that match the T5 vocabulary. In Prepare Input Representations step, token IDs are padded or truncated to a fixed length, and attention masks are created to ignore padding tokens during processing. In the step of generate encoder output, pass the tokenized input through the T5 encoder, then the encoder generates contextual embeddings for each token in the sequence. In aggregate term embeddings step, aggregate the term-level embeddings into a document vector for tasks requiring a single embedding vector (e.g., classification). Transformer Models rely on self-attention mechanisms, which are represented in Equation (6).

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (6)$$

Where, Q, K, and V are query, key, and value matrices, and d_k is the dimensionality of keys.In post-processing step, normalize the embeddings or perform dimensionality reduction (e.g., PCA) depending on downstream requirements.

G. Classification Models

The data samples were collected from the PAN 2017 competition Twitter dataset. 80% of the dataset was used for training and 20% of the dataset was used for testing while training the algorithm. Previously experimented with Traditional Neural Networks and some machine learning algorithms such as SVM and Random Forest with the hybrid approach. Finally hybrid model best suited to XGBoost classifier. XGBoost (Extreme Gradient Boosting) is used in the prediction of language variety. XGBoost is a tree-based ensemble machine learning algorithm known for its efficiency and high accuracy. Extract statistical features like **l**exical features, syntactic features, and stylistic features, BERT Embeddings**,** T5 Embeddings. Then, combine statistical features with BERT and T5 embeddings to produce the combined vector. Choose XGBoost classifier for training the combined vectors generated previously. More over All the researchers got better accuracy with XGBoost classifier as compared to number of well-known classifiers, which is the reason in selecting this classification model. Use cross-validation to assess classification model performance during training. Accuracy and F1-Score measures are used as Evaluation Metrics for presenting the experimental results of the proposed hybrid model.

1) Algorithm/pseudo-code

Input: PAN 2017 competition dataset
Output: Performance of language variety prediction
Step-1: Pre-process Data:
a. Clean and tokenize text.
b. Remove special characters and stopwords.
c. Perform Stemming/Lemmatization.
Step-2: Extract Features:
a. Extract Statistical Features:
- Lexical, syntactic, orthographic, stylistic features.
b. Generate BERT Embeddings:
- Tokenize text using WordPiece tokenizer.
- Extract contextualized embeddings from the BERT model. Apply pooling to obtain fixed-length document embeddings.
c. Generate T5 Embeddings:
- Tokenize text using SentencePiece tokenizer.
- Pass tokens through T5 encoder to generate embeddings.
- Aggregate embeddings into fixed-length vectors.
Step-3: Merge Features:
a. Normalize all feature sets (Statistical + BERT + T5).
b. Concatenate features into a unified document representation.
Step-4: Train XGBoost Classifier:
a. Split dataset into training and validation sets.
b. Train XGBoost using the merged features.
c. Perform hyperparameters tuning using cross-validation.
Step-5: Evaluate Performance:
a. Compute Accuracy, Precision, Recall, and F1 Score.

XGBoost Classifier uses gradient boosting to optimize predictions based on merged features. The objective function is represented in Equation (7).

$$obj = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i\right) + \sum_{k=1}^{K} \Omega\left(f_k\right) \quad (7)$$

Where, l is the loss function (e.g., logistic loss for classification), Ω regularizes the complexity of trees, and K is the number of trees. Gradient Boosting (XGBoost) iteratively improves predictions by minimizing the loss function, which is represented in Equation (8).

$$f(x) = \sum_{k=1}^{K} f_k(x) \quad (8)$$

Where, each $f_k(x)$ is a decision tree added sequentially to correct errors from previous iterations.

IV. Experimental Results

The experimental evaluation of native language prediction was conducted using various models, including a Traditional Neural Network (NN) with statistical features, BERT, T5, and hybrid approaches that combined BERT and T5 embeddings. The results demonstrate a clear advantage of transformer-based models over the statistical features with traditional NN, with BERT and T5 achieving significantly higher accuracy and F1 scores. The Table 2 presents a comparison of different models and their performance metrics like Accuracy and F1 Score on a native language prediction task.

TABLE II. The Performance of Different Models for Language Variety Prediction

Model	Accuracy	F1 Score
Statistical Features + Traditional NN	85.2%	83.7%
Statistical Features + BERT Embeddings with XGBoost Classifier	91.4%	91.5%
Statistical Features + T5 Embeddings with XGBoost Classifier	92.2%	92.8%
Average (Statistical Features, BERT, T5 Embeddings) with XGBoost Classifier	92.9%	93.0%
Merge (Statistical Features + BERT + T5 Embeddings) with XGBoost Classifier	**93.5%**	**93.4%**

Among the hybrid approaches, the appending of statistical features, BERT and T5 embeddings yielded the best performance, achieving an accuracy of 93.5% and an F1 score of 93.4%**,** surpassing both the standalone models and the averaged hybrid approach. These findings highlight the effectiveness of combining contextual and generative embeddings for capturing nuanced linguistic patterns, essential for distinguishing native language variations.

A. Discussion of Results

The baseline model is the conventional NN with statistical features, which achieves modest accuracy and F1-scores. Although it does a good job of capturing the fundamental patterns in the data, it is unable to model complex linguistic nuances such contextual dependencies and minute lexical or grammatical changes. When compared to transformer-based designs, traditional neural networks are insufficiently resilient for intricate applications such as language variety prediction. By using its bidirectional transformer architecture to produce contextualized embeddings, BERT with statistical features performs noticeably better than the conventional NN. This feature enables it to identify intricate textual patterns like regional grammar or colloquial expressions, which are influenced by variations in the local language. BERT is an effective stand-alone model for native language prediction because of its capacity to comprehend context at the token level.

Because of its text-to-text generative approach, which enables it to model linguistic dependencies more thoroughly, T5 with statistical features performs better than BERT. It is more successful at spotting minute dialectal or stylistic differences because it captures a better representation of text semantics. T5's architecture is ideal for creating and comprehending text representations, making it especially useful for tasks involving linguistic diversity. This model combines the advantages of both designs by averaging statistical features, T5, and BERT embeddings. The fact that T5's generative modeling and BERT's contextual knowledge complement each other is demonstrated by the increase in accuracy and F1 Score over solo models. Higher performance can be attained by hybrid strategies like embedding averaging, especially for tasks requiring sophisticated language comprehension. By keeping all of the contextual and generative information from both models, BERT and T5 embeddings add statistical features to produce a richer, more complete feature space. This method performs better at differentiating between native language variations, as seen by the greatest accuracy and F1 scores. The best approach is to add embeddings, which maximizes transformer-based models' advantages and produces cutting-edge results.

The ROC curve stands for the Receiver Operating Characteristic curve. It is a graphical representation of the performance of a binary classifier at different classification thresholds. The curve plots the possible True Positive rates (TPR) against the False Positive rates (FPR).

The **left side** of the curve corresponds to the more "confident" thresholds: a higher threshold leads to lower recall and fewer false positive errors. The extreme point is when both recall and FPR are 0. In this case, there are no correct detections but also no false ones. The **right side** of the curve represents the "less strict" scenarios when the threshold is low. Both recall and False Positive rates are higher, ultimately reaching 100%. If our model is correct in all the predictions, all the time, it means that graph if perfect. The ROC curve is represented in Fig. 2.

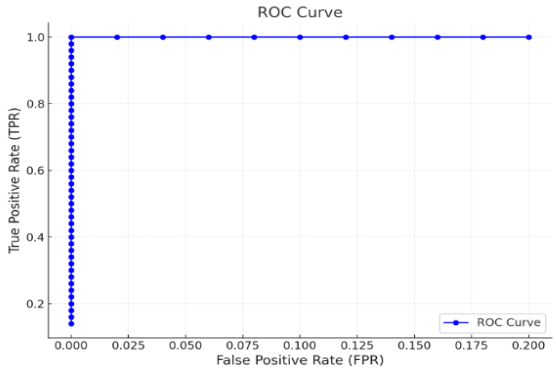

Fig. 2. ROC curve of the Hybrid Approach for Language Variety Prediction

The loss curve is plotted in Fig. 3.

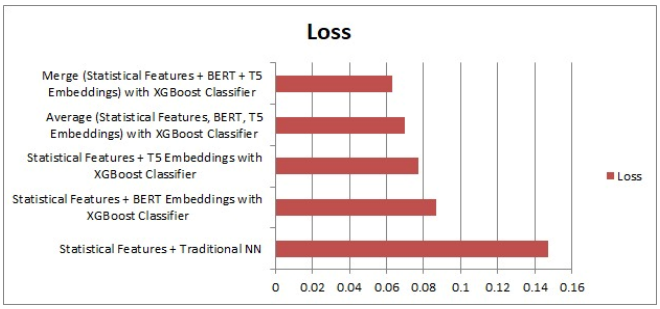

Fig. 3. Loss comparision of the Hybrid Approach for Language Variety Prediction

Accuracy Comparison is plotted in Fig. 4.

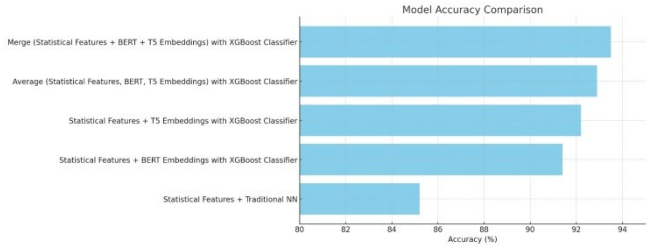

Fig. 4. Accuracies of the Hybrid Approach for Language Variety Prediction

Achieving high accuracy in a task like language variety prediction requires optimization at various stages, from feature extraction to model selection and evaluation. Based on the methodology and results described in the document, here are actionable strategies to improve accuracy further. Use a balanced and diverse dataset like PAN 2017, but supplement it with additional annotated datasets (e.g., tweets, articles) to improve the model's robustness to unseen variations. Ensure proper preprocessing techniques. Introduce data augmentation techniques to artificially expand the dataset. Use techniques like oversampling (e.g., SMOTE) or undersampling to ensure balanced training for each language variety. Explore more advanced statistical features, and more Transformer-Based Embeddings like BERT and T5. Fine-tune BERT and T5 on the specific dataset (e.g., PAN 2017) to adapt embeddings to the language variety prediction task. Combine multiple embedding sources like BERT and T5 embeddings with additional ones like Word2Vec, GloVe, or FastText. Use concatenation rather than averaging for richer feature representation.

Analyze misclassified samples to identify patterns and use insights from error analysis to refine feature engineering or retrain the model with augmented data.

V. Conclusion and Future Scope

The outcomes show how crucial sophisticated language models like BERT and T5 are for tasks requiring in-depth linguistic comprehension. By utilizing complimentary qualities, hybrid techniques further improve performance, and the best outcomes are obtained with the appended embedding strategy. These results demonstrate how transformer models can be used for high-stakes applications like identifying regional dialects or examining linguistic patterns in various datasets. In this work, the proposed hybrid model predicts language variety using statistical features, BERT, and T5 embeddings. For language variety prediction, the merging (appending) of statistical features, BERT embeddings, and T5 embeddings outperformed the averaging strategy (Accuracy: 92.9%, F1-Score: 93.0%) in terms of Accuracy (93.5%) and F1-Score (93.4%). The reason for this improvement is that by adding the embeddings rather than averaging them, a richer and more thorough representation is produced.

Future work will focus on merging BERT and T5 embeddings with other techniques and are planning to merge BERT and T5 embeddings with other embedding techniques such as Word2Vec, Glove, and fasttext embeddings for language variety prediction. We have also a plan of implementing BERT and CNN as classification models and BERT, T5 will be used for both embeddings as well as classification model for language variety prediction.

References

[1] Raghunadha Reddy T, Vishnu Vardhan B, Vijayapal Reddy P, “A Survey on Author Profiling Techniques”, International Journal of Applied Engineering Research, March 2016, Volume-11, Issue-5, pp. 3092-3102.

[2] https://pan.webis.de/clef17/pan17-web/author-profiling.html

[3] Para Upendar, T Murali Mohan, S. K. LokeshNaik, T Raghunadha Reddy, “A Novel Approach for Predicting Nativity Language of the Authors by Analyzing their Written Texts”, SPRINGER 6th International Conference on Innovations in Computer Science and Engineering, 17-18, August 2018, pp 17–22, Lecture Notes in Networks and Systems book series, volume 74.

[4] Dara Raju, T. Raghunadha Reddy, “Authorship Attribution using Content based Features and N-gram features”, International Journal of Engineering and Advanced Technology, Volume 9, Issue 1, October, 2019 pp. 1152 - 1156.

[5] Alexander Ogaltsov and Alexey Romanov, “Language Variety and Gender Classification for Author Profiling in PAN 2017”, Proceedings of CLEF 2017 Evaluation Labs, 2017.

[6] Raghunadha Reddy. T, P. Vijayapal Reddy, Nativity Language Prediction Using A Document Weighted Approach, International Journal of Creative Research Thoughts (IJCRT), Feb 2018, pp. 85-88

[7] Sameeah Noreen Hameed, Muhammad Adnan Ashraf, Qiao Ya-nan, Multi-Lingual Language Variety Identification using Conventional Deep Learning and Transfer Learning Approaches, The International Arab Journal of Information Technology, Vol. 19, No. 5, September 2022, pp. 705-712

[8] Daniel Escobar-Grisales, Juan Camilo Vásquez-Correa, Juan Rafael Orozco-Arroyave, Author Profiling in Informal and Formal Language Scenarios Via Transfer Learning, TecnoLogicas, vol. 24, num. 52, December 2021, pp. 1-25.

[9] Swapna M, Nikitha K, Contextually Propagated Term Weight based Approach for Author Profiling: Gender, Age and Language Variety Prediction J. Electrical Systems 20-5s (2024): 2834-2840

[10] Marcos Zampieri, Kai North, Tommi Jauhiainen, Mariano Felice, Neha Kumari, Nishant Nair, Yash Bangera, Language Variety Identification with True Labels, LREC-COLING 2024, pages 10100–10109, 20-25 May, 2024.

[11] Bassem Bsir, Nabil Khoufi, Mounir Zrigui, Prediction of Author’s Profile Basing on Fine-Tuning BERT Model, Informatica, Volume 48, 2024, pp. 69–78.

[12] Dominique Brunato, Andrea Cimino, Felice Dell’Orletta, Simonetta Montemagni, Giulia Venturi, Profiling–UD: a Tool for Linguistic Profiling of Texts, Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 7145–7151, Marseille, 11–16 May 2020

[13] Tommi Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, Krister Lindén, Automatic Language Identification in Texts: A Survey, Journal of Artificial Intelligence Research 65 (2019) 675–782.

[14] Chennam Chandrika Surya, Karunakar K, Murali Mohan T, R Prasanthi Kumari, Language Variety Prediction using Word Embeddings and Machine Leaning Algorithms, International Journal

for Research in Applied Science & Engineering Technology (IJRASET), vol. 10 Issue XII Dec 2022, PP.1616-1623

[15] Francisco Rangel, Paolo Rosso, Wajdi Zaghouani and Anis Charfi, Fine-grained analysis of language varieties and demographics, Natural Language Engineering , Volume 26 , Issue 6: Natural Language Processing for Similar Languages, Varieties, and Dialects , November 2020 , pp. 641 – 661

[16] T. Raghunadha Reddy, P. Vijaya Pal Reddy, P. Chandra Sekhar Reddy, "A New Supervised Term Weight Measure based Machine Learning Approach for Text Classification", International Conference on Intelligent Systems and Sustainable Computing, September 24-25, 2021, Intelligent Systems and Sustainable Computing, Smart Innovation, Systems and Technologies, pp 563–571, vol 289.

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding", Proc. NAACL-HLT, 2019, pp. 4171–4186.

[18] Swathi Ch, Karunakar K, Archana G, T. Raghunadha Reddy, "A New Term Weight Measure for Gender Prediction in Author Profiling", Proceedings in Advances in Intelligent Systems and Computing, Volume 695, PP. 11-18, 2018. Intelligent Engineering Informatics.

[19] Zaki, Muhammad Zayyanu, Revolutionising Translation Technology: A Comparative Study Of Variant Transformer Models - Bert, Gpt And T5, Computer Science & Engineering: An International Journal (CSEIJ), Vol 14, No 3, June 2024