LING 380: English Dialect Classification

Daniel Wang
Yale University,
Statistics & Data Science
daniel.wang.dyw8@yale.edu

Jewon Im
Yale University,
Computer Science
jewon.im@yale.edu

Michael Toben
Yale University,
Statistics & Data Science
michael.toben@yale.edu

Abstract

Dialect classification is a subset of language classification in machine learning and focuses on distinguishing linguisitic variations within the same language. In natural language processing (NLP), dialect classification is an important task that is crucial for serving a diverse user base effectively by leveling linguistics equity and reducing bias for a "standard" dialect. This study evaluates the performance of neural network models for English dialect classification, testing the fine-tuning of two BERT models, DistilBERT and TinyBERT, and comparing the performance to a baseline Naive Bayes model. Using the English corpora, specifically the Global Web-Based English (GloWbE) corpus in addition to testing datasets, we trained these models on 20 English dialects from different countries. Our results reveal that training data diversity has a profound impact on model performance, with underrepresented dialects frequently being misclassified. In addition, we find that model complexity contributes to performance but still requires further tuning and data representation for success. Overall, our work sheds light on the potential of neural network for dialect classification and insights into their limitations.

1 Introduction

Natural language processing (NLP) methods have positively influenced and democratized technology and information. One such NLP task is language classification, which is used in all sorts of everyday media such as translation apps (e.g. Google Translate), virtual chatbots, search engines, and academic research. Clearly, optimizing NLP methods will have a clear impact on the safety and efficacy of everyday technologies. While roughly 7200 languages are spoken in the world every day (Anderson, 2004), there are even more dialects spoken. Dialects are defined as varieties of certain languages spoken by particular groups of people. Hence, even languages as commonly spoken as English can have more than 100 dialects! Training accurate models for dialect classification will yield similar benefits as language classification, yet few models have been optimized to do so. This is likely due to 1) lack of training data in less-spoken dialects and 2) model performance may not be very accurate because of the difficulty in distinguishing between similar dialects. However, we believe that with the success of BERT-trained models in other classification texts, neural networks are a solid option for dialect classification. This paper is designed to answer the questions: how do large language models recognize dialectal variations within the same language (e.g. English), and more specifically, can a language identification neural network model also be trained to recognize and accurately identify per-country dialectal variations in English?

The first large-language models developed, such as BERT, have been optimized for tasks such as language classification. Since then, more complex models have been developed to perform even better at such tasks while retaining computational efficiency. However, a similar but often overlooked task in the field of natural language processing is dialect classification.

Dialect classification is important because building language models that are successful in distinguishing between dialects can lead to better linguistic equity, such as reducing bias towards one dialect or flagging others as less correct. Refining dialect classification models will lead to language models being more useful to those who may not speak the most commonly spoken dialect of a language.

In a historical sense, refining dialect classification may also allow researchers to unearth the nationalities, histories, and origins of unattributed text, aiding research in humanities fields as well.

Although dialect classification is similar to language classification, dialectal differences tend to

be much more subtle than differences between languages. For example, in English, all twenty-six letters in the alphabet are used across dialects, and the majority of words are similar between dialects. Oftentimes, the difference between dialects comes down to small differences in expression, syntax, and proper nouns. In other words, dialect classification is much more difficult than language classification because the differences between different dialects of the same language tend to be much smaller than the differences between languages. For example, when given chunks of text in different English dialects to classify, a model may have to look at syntactic subtleties, whereas a language classifier can use more obvious features such as word order and writing system to differentiate between languages.

Models trained for language classification oftentimes cannot generalize to dialect classification because of the nuances required to differentiate between dialects. These nuances may not be picked up by models trained specifically for language classification, where the task involves character and word identification/recognition with more obvious differences in syntax. Furthermore, dialects can be subtle within the same language depending on variations in spelling, vocabulary, syntax, and cultural nuances, making it difficult for NLP models trained on standard languages to generalize to dialect classification.

Finally, high-quality data is much harder to find for dialect classification than language classification. Many texts exhibit a range of dialects of the same language; however, they may only contain one unique language. For example, a text written in Standard American English may include a cultural reference to British English (e.g. A quote from BBC, or a character who speaks in British English). This mixing reduces the quality of the data and has implications on both training and testing. During training, this mixed data model may incorrectly learn features of one dialect as another and during testing, this mixed data may create uncertainty within the model's predictions.

Some past papers have done work specifically with using deep learning to classify dialects based on audio recordings. While the models incorporating audio recording must be complex and multimodal, dialects have audible differences that cannot be purely captured within the text, making it a viable method. For instance, in Chinese, Mandarin and Cantonese use the same characters, yet have different pronunciations. Audio recordings are the most clear way to distinguish between words in both cases. While the differences in pronunciations among English dialects are less obvious, they are still present. For example, the word either is pronounced eye-thuh in British English but ee-thuhr in American English.

In particular, we find the BERT model fascinating not only because of its common usage in language classification, a task that closely mirrors dialect classification, but also its unique architecture which incorporates bidirectionality. BERT has been pre-trained on the entire English Wikipedia, meaning it has likely been trained on text from all dialects (Devlin et al., 2019). Therefore, we hypothesize that though not fine-tuned for the specific task of dialect classification, BERT may have been optimized to observe certain distinguishing features of dialects. Finally, BERT has several different variants, two of which we will be investigating: TinyBERT and DistilBERT. According to Jiao et al., TinyBERT is a four-layer neural network that is the most compact BERT model, while according to Sanh et al., DistilBERT has 4-6 layers (Jiao et al., 2020, Sanh et al., 2020). Both are more compact than the original BERT model, which allows us to test computational efficiency.

All in all, in this paper, we will evaluate how well BERT models can perform dialect classification after being fine-tuned to do so. We will compare the performance of BERT models versus a Naive Bayes model that utilizes Bag-of-Words, which is just a mere counter of words, to determine if models are actually able to pick up differences between dialects or if they do worse than a model that just counts the number of words per text. By doing so, we hope to answer our big-picture question: How do large language models recognize dialectal variations within the same language (e.g. English)? More specifically, can a language identification neural network model also be trained to recognize and accurately identify per-country dialectal variations in English?

2 Background

2.1 Naive Bayes and Natural Language Processing

Probabilistic models have long been used for natural language processing for as long as people have been interested in analyzing text. The first instance of text analysis was in 1976, when Stephen E. Robertson and Karen Jones published "Relevance

weighting of search terms," which utilized similar counting model to Naive Bayes (Robertson et al., 1976).

The Naive Bayes started to become more widely used in the 1990s and 200s. At that time, it was often used for spam filtering. Today, Naive Bayes is still used often, especially on identification, classification, and sentiment analysis tasks. It is known to be computationally efficient simple, and effective, but it is typically more effective with smaller datasets.

2.2 BERT

BERT (Bidirectional encoder representations from transformers) was introduced in 2018 by Google. Though it uses encoder-only architecture, its aim was to learn vector embeddings, which set off the AI craze as one of the first language models. BERT was originally trained with all of the English Wikipedia data, with two sizes: BERTBase, at 110 million parameters, and BERTLarge at 340 million parameters (Devlin et al., 2019). Two years later, other smaller models, including TinyBERT were released (with only 4 million parameters)(Jiao et al., 2020).

BERT has four layers; first, a tokenizer layer that converts data into a sequence of integers known as tokens. Then, it feeds the sequence to an embedding layer that represents each sequence with a vector. Then, it utilizes an encoder which consists of transformer blocks that utilize self-attention. Finally, its last layer, the task head, typically uses a softmax function to convert vectors into a probability distribution. For transfer and fine-tuning, the task head is replaced with a task-replacement module (Devlin et al., 2019).

2.3 Language Classification

The history of language classification starts roughly around the same time as text analysis started to gain popularity. There were two schools of thought in this field: symbolic (rule-based) and stochastic (statistical). After Roberston et al., several other influential papers in language classification include Collobert et al. (2011), which was one of the first papers to encourage using neural networks for NLP tasks. Then, Kim (2014) introduced the idea of applying CNN specifically to classification tasks. Word embeddings were also improved, such as with the introduction of ElMo in Peters et al. (2018).

2.4 Early Dialect Classification

Early dialect classification methods were often reliant on programming linguistic features such as phonology, morphology, and syntax into the models. Later on, more automated statistical techniques were used, such as support vector machines (SVN) and decision trees, and eventually deep learning which dominates the modern landscape. Recent advancements in deep learning have led to the development of more sophisticated models for dialect classification. Literature evaluating CNNs and RNNs have shown great promise with deep learning methods (Yin et al.).

With the availability of tools like HuggingFace combined with easy access to data, transfer learning approaches, such as fine-tuning pre-trained language models like BERT, have become increasingly popular for dialect classification tasks.

With the rise of multi-modal models, acoustic features have been included in the datasets as well, with the development of techniques like Single Frequency Filtering (SFF) and Zero-Time Windowing (ZTW) that are able to capture dialectal variation in pitch and tone variation (Kethireddy et al.).

2.5 Modern Dialect Classification

Today, the field of dialect classification is inhibited by two issues: first, there is not much data in corpora to train lesser-known dialects. In the field, some techniques are being explored to mitigate this issue by training models on similar dialects and specifically identifying features that distinguish differences, but this approach is linguistics-heavy and rare. Others have been developing evaluation metrics other than traditional accuracy metrics to factor in the lack of quantity of test and training data for rare dialects compared to those spoken and written more frequently.

Several papers that are relevant to our project have explored dialect classification with deep learning. Najafian et. al. (2016) addresses the challenges of acoustic modeling for Automatic Speech Recognition (ASR), particularly in recognizing regional accents. While this paper is more about acoustic modeling, we found some of its insights particularly applicable towards our task. It highlights how DNN-based models can benefit from multi-accent learning strategies, even with small amounts of data. This approach, which enhances model robustness by supplementing training data with targeted examples, is directly applicable to

our project. Based on the weakness mentioned in the prior paragraph of potentially using language identification models for dialect identification tasks, we may include training examples of phrases/sentences that have the same words but in a slightly different order as representative of different dialects to help the model improve performance.

Dunn (2019) models syntactic variation across 14 national varieties of English, using grammar induction and text classification with the goal of dialectical identification. It introduces data-driven language mapping to select relevant dialects. The key insight of this paper that we think is beneficial for our project is that the model is improved by diversifying the medium of the training data. Specifically, Dunn's model was more accurate at dialect identification in non-traditional sources (e.g. Short social media posts) if these were included in the training data. While this may sound trivial, it highlights that the unique qualities of dialects models pick up on may not be transferable to different sources of test data.

Datta (2023) investigates how large language models interpret dialectal variations in English with the context of news and social media. It shows that using dialectal information in the model improves its performance, especially when fine-tuning pre-trained models like BERT, BART, and T5. These findings are relevant to our project as they demonstrate that models can effectively distinguish dialects such as American and British English when trained on diverse, dialect-specific data.

Through our research design, we hope to not only replicate some of the previous work done with neural networks on linguistics, but we hope to make judgements and conclusions about the success of using deep learning techniques on dialect classification ourselves to answer the linguistic question of what truly differentiates dialects in English, and whether neural network models are sophisticated enough to parse these differences.

3 Methods

3.1 Models and Network Architecture

To evaluate performance on the task of English dialect identification, we compared several language models, including a baseline Naive Bayes model and two pre-trained transformer-based models: DistilBERT and TinyBERT.

BERT (Bidirectional Encoder Representations from Transformers) is a language representation model that uses a bidirectional transformer architecture to capture contextual relationships between words in a sentence. Unlike traditional left-to-right models, BERT simultaneously considers both left and right contexts, allowing it to generate richer word embeddings. BERT has proven particularly effective for token classification, text classification, and sequence labeling tasks, which makes it especially suitable for tasks like dialect identification where subtle contextual cues can differentiate dialects (Devlin et al., 2019).

We chose to evaluate BERT models for dialect identification because of their ability to handle nuanced variations in language usage, such as differences in vocabulary, grammar, and word choice, that are often indicative of specific dialects. These nuanced features can be challenging for simpler models, such as Naive Bayes, to identify.

3.2 BERT

In terms of how DistilBERT and TinyBERT compare to one another, both DistilBERT and TinyBERT are compact versions of BERT, designed to reduce computational overhead while retaining high performance. DistilBERT is a distilled version of BERT, created through knowledge distillation, where a smaller model is trained to mimic the behavior of the larger BERT model. It has 6 layers (half the depth of BERT-Base) and achieves roughly 97% of BERT's performance while being 60% faster (Sanh et al., 2020). TinyBERT is even more compact, typically containing 4 or 6 layers. It employs a two-stage knowledge distillation process: general distillation on a large corpus followed by task-specific distillation on downstream tasks (Sanh et al., 2020). TinyBERT is optimized for environments with limited computational resources, achieving performance comparable to BERT while being significantly smaller and faster.

3.3 Naive Bayes Model

We included a baseline Naive Bayes model to contrast the performance of modern transformer-based models with a traditional, interpretable approach. Naive Bayes is a probabilistic classifier that assumes feature independence and is often used for text classification tasks due to its simplicity and efficiency (Sharma et al, 2023). Comparing these models allows us to measure the added value of transformers for dialect identification.

3.4 Training

We utilized the Global Web-Based English (GloWbE) corpus as our primary training dataset. GloWbE contains 1.8 billion words spanning 20 dialects of English, providing a diverse and robust resource for model fine-tuning. We used a sample dataset from GloWbE spanning a total of 2.1 million words across the dialects. The dialects include:

Code	Dialect
AU	Australian English
BD	Bangladeshi English
CA	Canadian English
GB	British English (Great Britain)
GH	Ghanaian English
HK	Hong Kong English
IE	Irish English
IN	Indian English
JM	Jamaican English
KE	Kenyan English
LK	Sri Lankan English
MY	Malaysian English
NG	Nigerian English
NZ	New Zealand English
PH	Philippine English
PK	Pakistani English
SG	Singaporean English
TZ	Tanzanian English
US	American English
ZA	South African English

Table 1: Country codes and their corresponding English dialects.

The pre-trained DistilBERT and TinyBERT models were sourced from Hugging Face and fine-tuned on the GloWbE corpus for the dialect identification task. We further fine-tuned models by training the models to predict one of the 20 dialect labels based on input text. We also worked on adjusting hyperparameters, including learning rate, batch size, and epochs, to optimize performance.

For the Naive Bayes baseline, we implemented a simple bag-of-words feature extraction pipeline combined with the Multinomial Naive Bayes classifier.

3.5 Testing

To evaluate the performance and generalizability of our fine-tuned models, we expanded the dataset beyond the original movie (1.6 mw) and TV show (2.1 mw) data sourced from the English Corpus. Additional data were incorporated from two major sources within the English Corpus: the News on the Web (NOW) corpus (1.7 mw) and articles related to the COVID-19 pandemic (3.2 mw). These sources were selected for their breadth and diversity, providing contemporary examples of English usage across various global dialects. By including these newer datasets, we aimed to assess the models' ability to generalize beyond the more structured and domain-specific nature of the initial training data.

The raw text data from the NOW and COVID corpora required significant preprocessing to make it suitable for machine learning evaluation. Each text file contained unstructured entries with varying levels of metadata. We first identified and extracted country-specific information by associating each text entry with its respective country of origin, as indicated in the available metadata. To ensure consistency, we filtered for texts where English was the primary language. Non-English entries and texts with multiple listed languages were excluded, reducing potential noise in the evaluation dataset. Country names were converted into the conventional two-letter code system that we established earlier (Table 1).

The resulting test dataset represents a diverse collection of modern-day English dialects, spanning web-based journalism, COVID-related news articles, and media sources (Table 2). By introducing this external data, we sought to evaluate the robustness of the fine-tuned models when exposed to real-world linguistic variation beyond the GloWbE corpus. This step was crucial for assessing the models' ability to recognize and differentiate dialectal features in contexts that reflect contemporary usage and evolving patterns of English.

Model performance was evaluated using accuracy as the primary metric, though we also utilized F1 Score and Precision Score. Specifically, we compared the accuracy of the fine-tuned transformer models on dialect identification and the accuracy of the baseline Naive Bayes model, as well as the reported accuracy of pre-trained BERT models on standard language identification tasks (sourced from Hugging Face benchmarks).

To further understand our results, we visualized our model's accuracy with a confusion matrix for us to parse whether there were dialects that the models performed surprisingly well in or surprisingly poorly in.

On a task basis, success was defined as the model correctly identifying the dialect of the input text.

Country	GloWbE	Media	COVID	NOW	Total
AU	130	19	221	147	517
BD	35	0	35	21	91
CA	117	60	281	356	814
GB	339	125	384	402	1250
GH	34	0	35	61	130
HK	35	2	30	6	73
IE	88	2	198	278	566
IN	85	0	342	388	815
JM	35	0	15	15	65
KE	38	0	82	33	153
LK	42	0	33	16	91
MY	38	0	100	73	211
NG	38	0	172	143	353
NZ	72	0	115	126	313
PH	38	1	99	118	256
PK	45	0	63	80	188
SG	39	0	141	98	278
TZ	31	0	11	6	48
US	339	592	1815	396	3142
ZA	40	4	190	197	431

Table 2: Summary of data sources for each country across GloWbE, Media, COVID, and NOW corpora, including total counts.

Overall, we aimed for the fine-tuned models to achieve an accuracy rate for dialect identification comparable to their accuracy on language identification tasks. In addition, we experimented with various hyperparameter settings to maximize identification accuracy, ensuring a rigorous evaluation of model performance.

4 Results

Overall, for the Media Accuracy, both DistilBert and TinyBert had equivalent accuracies, with Naive Bayes having very poor accuracy (Table 3). For the COVID dataset, Naive Bayes performed better, but Distilbert was still the best performing model, followed by TinyBERT, which had the second best performing model. On the NOW dataset, Distilbert still performed the best, followed by Naive Bayes. TinyBERT performed the worse. Overall, both BERT models' accuracy decreased from media dataset to the COVID to NOW accuracies.

4.1 Results by Dataset

For the Naive Bayes model, when looking at any single testing set, the accuracy, F1 Score, and precision score are similar. The similarity of these scores indicates that our accuracy was a robust evaluation metric. We have similar findings for DistilBert and TinyBERT, but it is worth noting that TinyBERT had the steepest dropoff between precision and accuracy. This is attributed to the fact that TinyBERT predicted US for all datapoints except for 1 (GB, which it did so correctly), which diminishes the appeared strength of TinyBERT.

Looking at the COVID dataset, the Naive Bayes' model also had similar accuracy, F1 Scores, and Precision. Distilbert had a greater gap between its accuracy score and precision and TinyBERT had a huge decrease between accuracy and precision score. It is worth noting that the COVID dataset has more data from lesser-known dialects than the media dataset, implying that the BERT models are stellar at predicting popular dialects that the majority of their training data comes from, but less so for less popular dialects. The Naive Bayes model, however, retains its consistent accuracy and precision score marks.

Finally, for the NOW dataset, the most diverse test dataset, the patterns listed above are much more prevalent. As this dataset has the greatest proportion of data not from the most popular dialects, we recognize not only the lowest accuracies for the BERT models, but also the greatest differences

Model	Media Accuracy	COVID Accuracy	NOW Accuracy	Overall
Naive Bayes	0.02	0.39	0.21	0.30
DistilBERT	0.76	0.51	0.38	0.42
TinyBERT	0.76	0.42	0.13	0.16

Table 3: Accuracy of different models across datasets.

Model	Accuracy	F1	Precision
Naive Bayes	0.02	0.00	0.01
DistilBERT	0.76	0.77	0.78
TinyBERT	0.76	0.66	0.58

Table 4: Performance metrics for different models on the Media dataset.

Model	Accuracy	F1	Precision
Naive Bayes	0.39	0.34	0.37
DistilBERT	0.51	0.44	0.41
TinyBERT	0.42	0.25	0.17

Table 5: Performance metrics for different models on the COVID dataset.

between their accuracies and F1 scores.

Overall, DistilBERT demonstrated the most consistent and robust performance across all datasets and countries due to its high accuracy, F1 Score, and precision. Despite the drop off between accuracy and precision in the now dataset, it still ranks the highest amongst all three models.

4.2 Results by Country

Looking at data by country, all models were the most successful at predicting American English data correctly (apart from Naive Bayes for media). This makes sense as in the training data, American English data made up the most cases.

As TinyBERT predicted US for nearly every datapoint in every dataset, and received 0 accuracy for all countries in the other datasets. However, at further glance, DistilBERT and Naive Bayes also produced similar results–GB was the most accurate for both of the two (and GB was the second most prevalent country in the training corpus), and so on.

4.3 Validation Accuracy

Regarding performance on training data, we also calculated validation accuracies for all 3 models. DistilBERT had the highest validation accuracy, followed by Naive Bayes, and then the TinyBERT. The pattern of DistilBERT being the highest with

Model	Accuracy	F1	Precision
Naive Bayes	0.21	0.18	0.22
DistilBERT	0.38	0.29	0.26
TinyBERT	0.13	0.03	0.02

Table 6: Performance metrics for different models on the Now dataset.

Dialect	NB	DistilBERT	TinyBERT
GB	0.01	0.61	0.01
US	0.00	0.85	1.00
CA	0.00	0.18	0.00
AU	1.00	0.15	0.00
IE	0.00	0.00	0.00
PH	0.00	0.00	0.00

Table 7: Performance metrics for different models on the Media dataset by country of dialect.

Naive Bayes and TinyBERT performing poorly being consistent with validation data proves that metrics for other tests are robust. Simultaneously, the validation data being less accurate for all models than some test datasets the media and COVID datasets points to potential underfitting.

5 Discussion

An interesting point to note is that TinyBERT showed high performance on the Media dataset, particularly for US data, but struggled significantly with the Now dataset and non-US data in other datasets. This is a result of TinyBERT nearly predicting US for all data points, which either means either TinyBERT overfit on its training data that was heavily US-centric, or TinyBERT was not computationally complex enough to grasp the nuances of each dialect. DistilBERT's triumph over TinyBERT underscores that though DistilBERT being a compressed form of BERT performed well, further compressions may lead to severe underperformance. Despite Naive Bayes' poor performance, it still had good accuracy in the COVID dataset and NOW dataset, performing particularly well in Great Britain and American English sources. As the Naive Bayes' model is just a bag-of-words model,

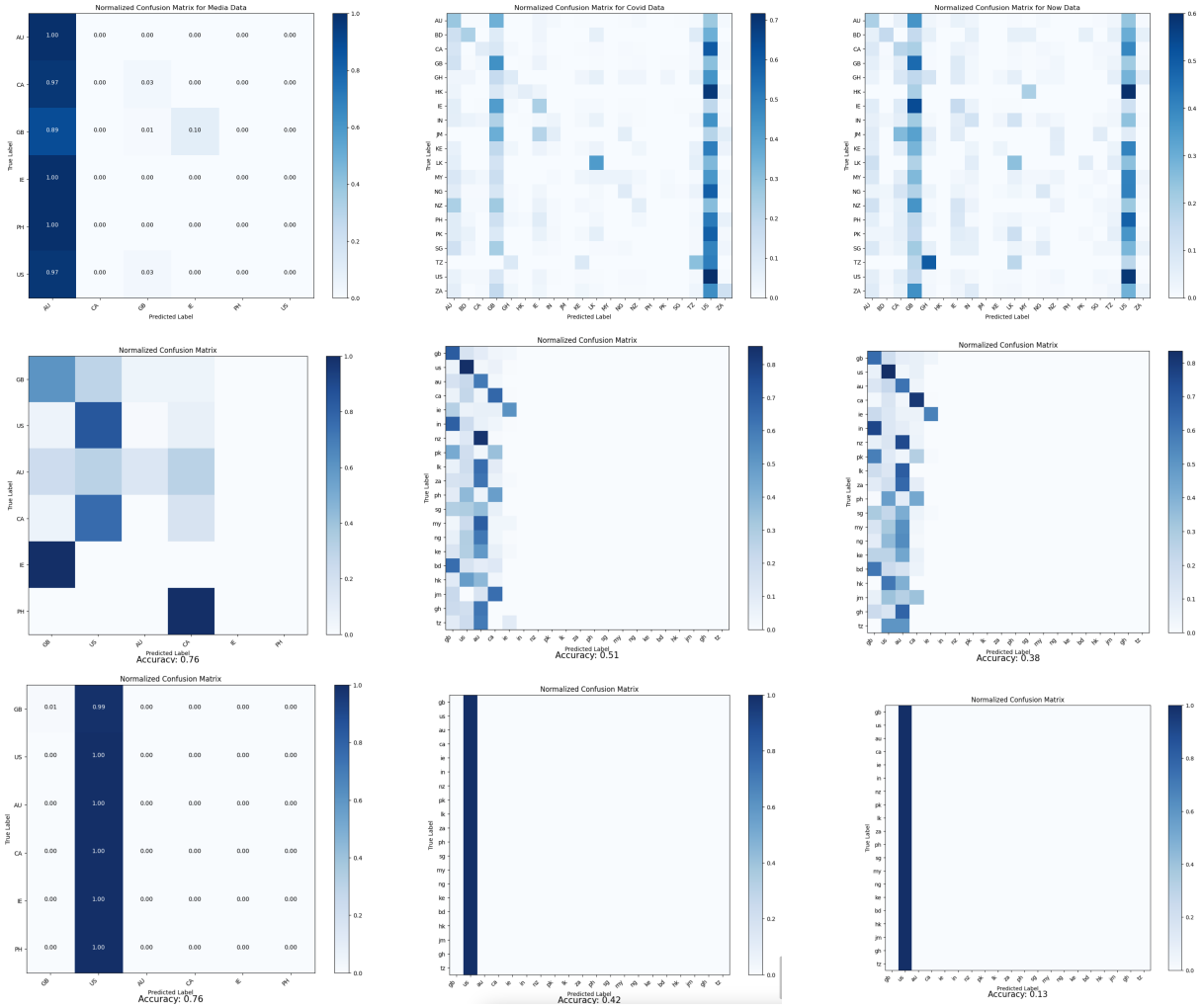

Figure 1: A grid of confusion matrices showing model performance across datasets. From top to bottom: Naive Bayes, DilstilBERT, and TinyBERT models. From left to right: Media, COVID, and NOW datasets.

Dialect	NB	DistilBERT	TinyBERT
GB	0.45	0.71	0.00
US	0.72	0.86	1.00
CA	0.08	0.67	0.00
AU	0.27	0.60	0.00
IE	0.24	0.53	0.00

Table 8: Performance metrics for different models on the COVID dataset by country of dialect.

Dialect	NB	DistilBERT	TinyBERT
GB	0.46	0.63	0.00
US	0.58	0.84	1.00
CA	0.23	0.80	0.00
AU	0.17	0.61	0.00
IE	0.15	0.57	0.00

Table 9: Performance metrics for different models on the NOW dataset by country of dialect.

pockets of Naive Bayes' strength underscore that while word count is not everything there is to a dialect, it is not nothing either. Naive Bayes was also the only model to perform strongest during COVID. We attribute this to the usage of proper nouns in COVID (likely world leaders, places, healthcare systems, etc.) that lead to more differences between countries. Such a performance also underscores the importance of using different models to classify different tasks.

These results also highlight an issue stated in the background section of this paper, which is that the models' performance varied significantly across countries, with generally better results for US data. As American English was the most dialect in the training data, it made sense that all models were best at predicting American English.

Our results also highlight the trade-offs between

model complexity, performance, and generalizability. While DistilBERT performed the best, it did require more computational resources than TinyBERT and Naive Bayes. On the other hand, Naive Bayes was in many ways not necessarily a worse model than TinyBERT despite being much less complex.

5.1 Robustness Checks

When doing our analysis, we also took note of possible confounding variables that could have accounted for variability within the models. For instance, one might state that the BERT models were not actually learning differences between dialects, just using context/performing a similar task to the Naive Bayes. Showing that at least DistilBERT performs better than Naive Bayes and that TinyBERT performs much worse proves that this is not the case. Another potential confounding variable we considered was that accuracy may have been impacted by how close the test data was to our train data. To resolve this, we decided to compare validation accuracies, which were lower than media and COVID for all models but higher than NOW. Such results point to prevalence of popular dialects such as US, GB being a greater indicator for model performance on a dataset rather than the text itself.

5.2 Summary

Overall, our analysis reveals the strengths and weaknesses of Naive Bayes, DistilBERT, and TinyBERT across different datasets and English-speaking countries. DistilBERT emerges as the most reliable and consistent performer, suggesting that BERT models indeed can distinguish between tasks. However, the underperformance from TinyBERT and pockets of strength from Naive Bayes highlight that computational complexity is not the end-all-be-all.

The variations in performance across countries and tasks highlights the issue of inequity in training data. At the end of the day, all models are limited by the training data they are exposed to. To further increase dialect classification accuracy, regardless of model complexity, the issue of diversity in the corpus must be addressed.

6 Conclusion

All in all, our designs and experiments have yielded answers to our two guiding questions: 1) How do large language models recognize dialectal variations within the same language (e.g. English)? and 2) More specifically, can a language identification neural network model also be trained to recognize and accurately identify per-country dialectal variations in English?

To answer question 1, our results indicate that some BERT models, such as DistilBERT do learn details about each dialect apart from word count, as it performs better than a Naive Bayes model that uses Bag-of-Words. Simultaneously, our results indicate training data bias has a very big impact on model performance for certain languages. Likely due to how similar dialects are, BERT, TinyBERT, and Naive Bayes will both resort to predicting the most common dialects for less common dialects.

To answer question 2, the overperformance of DistilBERT does indicate that neural networks can be trained to recognize and accurately identify per-country dialectal variations in English. As stated previously, the extent of the accuracy heavily depends on the training data.

Our conclusions point us into new directions with new questions and new focuses. On a smaller scale, first, as Professor Frank informed us in class, there is still the possibility that BERT models might be using word embeddings to generalize to words that are poorly represented in the training data. A solution to this could be to examine the embedding space to identify clusters or patterns that indicate overgeneralization. Additionally, generating more training data with synthetic examples or underrepresented words might help improve the model's robustness. Fine-tuning BERT on a domain-specific corpus or leveraging adversarial training to test its resilience against edge cases could also address these concerns. Next, we are interested in testing out more BERT models. There was a huge discrepancy between DistilBERT and TinyBERT. Would it be true that even larger models perform much better?

On a larger scale, our research provides great evidence for 'poverty of the stimulus' in language modeling, especially as it pertains to tasks like dialect classification. The steep drop-offs in TinyBERT and DistilBERT between test/validation data with more common dialects versus less common dialects is statistically significant. It emphasizes that to improve dialect classification, rather than focusing on model complexity itself, a great direction is to find more sources or generate more sources of uncommon dialects.

References

Stephen R. Anderson. How many languages are there in the world?

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237.

Delip Rao. 2023. Understanding naive bayes for natural language processing (nlp). *ResearchGate*.

S. E. Robertson and K. Sparck Jones. 2016. Relevance weighting of search terms. *ResearchGate*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Suwon Shon, Ahmed Ali Ahmed, Yue Meng, Tara Hebert, Shinji Watanabe, and Stefan Hahn. 2022. Deep neural architectures for dialect classification. *The Journal of the Acoustical Society of America*, 151(2):1077–1089.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1702.01923*.

Yichun Yin, Cheng Chen, Lifeng Wan, Weiping Gao, Zhonghua Liu, Haolin Wang, and Fuli Wang. 2020. Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351*.

A Appendix

All code can be found the GitHub repository, dialect-classifier.