Predicting Tweet Locations:
A Machine Learning Approach to Text Analysis

Ramanan Srirajan (rsrir2), Bochuan Zhang (bochuan3), Vladislav Fedorov (vvf2)

STAT 432: Basics of Statistical Learning

1. Abstract

The rise of social media platforms like Twitter has brought a significant increase in the availability of real-time data. However, location information is often incomplete or unreliable, which limits its potential for applications such as disaster response or public health monitoring. This project explores three models for inferring tweet locations based on text: Random Forests, Deep Learning with BiLSTM, and Logistic Regression. The Disaster Tweets dataset served as the foundation, and our focus was on handling issues such as missing and invalid locations, severe class imbalances, and language variability. Each model has strengths and trade-offs. Random Forests excelled at capturing regional word patterns but needed more depth, while sentiment analysis with TF-IDF captured emotional and contextual nuances in the texts. BiLSTM handled complex patterns but faced overfitting and imbalance issues. Despite these challenges, our models achieved serviceable results, showing the potential for using textual content to infer geographical locations. Logistic Regression was efficient and interpretable but struggled with precision-recall trade-offs.

2. Literature

2.1 Background and Foundational Methods

Geolocation prediction, which uses textual content and metadata to infer the location of social media posts, has been widely studied. Techniques like TF-IDF help identify regional text patterns, while sentiment analysis leverages emotional tone to improve accuracy. Logistic Regression and Random Forests are commonly used; the former is simple and interpretable but struggles with imbalanced datasets, while the latter handles high-dimensional data effectively but is computationally intensive (Georganos et al., 2019). Challenges such as misleading data from posts about other locations can be mitigated through inference techniques (Lamsal & Harwood, 2022).

2.2 Applications of Traditional and Deep Learning Models

Deep learning models, including BiLSTM and CNN-BiLSTM, excel at capturing complex patterns but often require large datasets and face overfitting risks (Sharma et al., 2021). Transformer-based models like BERT achieve high precision by combining textual and metadata features but are computationally expensive (Lutsai & Lampert, 2023). Hybrid approaches integrating text, metadata, and social networks further improve prediction accuracy (Bakerman et al., 2018).

2.3 Current Challenges

Persistent challenges include data imbalance, urban area bias, and vague textual references. Solutions such as location-indicative words (LIWs), stratified sampling, and lightweight models have shown promise, with LIWs reducing median errors by over 200 km on regional datasets. Building on these methods, this study employs Random Forests, Logistic Regression, and attention-enhanced BiLSTMs, along with clustering techniques, to improve prediction accuracy in underrepresented regions while addressing scalability and data imbalance issues.

3. Data

The Disaster Tweets dataset from Kaggle served as the foundation for our analysis. This dataset includes textual content from tweets and associated locations. The preprocessing phase focused on handling missing data, cleaning invalid entries, and reducing the abundance of location classes. The pertinent columns of the original dataset are location (where the tweet was posted) and text (tweet content).

3.1 Additional Datasets Used:

- World Cities Database: This database includes details such as city names, corresponding countries, and geographical coordinates. SimpleMaps
- U.S. Cities Database: This dataset provides city names, state affiliations, and population data for cities in the United States. SimpleMaps
- Country Mapping - ISO, Continent, Region: This dataset maps countries to their respective ISO codes, continents, and regions. Kaggle

3.2 Challenges

- Missing Location Data: A significant number of tweets were missing valid location information. To ensure the models were trained on reliable data, we disregarded entries without any location entries during the training phase. While this approach reduced the size of the training dataset, it eliminated noise that would otherwise have skewed the results.
- Invalid Locations: The dataset contained entries with invalid or non-geographical locations ("Mars" or "Nowhere"). These entries were identified and removed using additional datasets for reference, including world and U.S. cities, to retain only “real” location values.
- Overly Specific Locations: Many tweets included highly specific locations, such as small towns or neighborhoods, which added unnecessary complexity to the classification task. To address this, we grouped these granular locations into broader geographical regions. For instance, specific cities were mapped to their corresponding countries or regions (for example "Paris" to "Europe").
- Class Imbalances: Certain regions, such as North America, were overrepresented in the dataset, while others, such as Africa, had far fewer entries. During the data processing, we did not take any actions to fix this, but we addressed it in the modeling process which will be talked about later.

3.3 Data Cleaning

- Removing Duplicates: Duplicate tweets were identified and removed to prevent the model from being biased toward repeated entries.
- Language Filtering: Since the focus of this study was on analyzing English-language text, non-English tweets were excluded using automated language detection tools.

3.4 Region Assignment

The raw dataset included highly specific locations (small towns, neighborhoods), ambiguous entries, and invalid or fictional locations. To address these issues, we used an iterative process to map each location to broader regions for the sake of practicality and better class balance.

3.4.1 Mapping Locations to Countries

We started by identifying valid city names using the Worldcitites and Uscities datasets. These datasets provided lists of cities worldwide and in the United States, along with their corresponding countries and states respectively. For each tweet:

- If the location contained or matched an entry in Uscities's cities or states, it was assigned to the United States.
- Otherwise, if the location matched an entry in Worldcities, it was mapped to the respective country listed.
- Locations that didn't match any entries were labeled invalid and removed.

This process eliminated incorrect non-geographical locations, reducing noise and ensuring the remaining data was valid and usable.

3.4.2 Experimenting with Country-Based Classification

After that, we were going to classify tweets at the country level, and while this method offered more granularity, it introduced other issues:

- Popular countries like the United States and India had a large number of tweets, while many others, like Latvia, had very few. This imbalance made it difficult for models to generalize effectively.
- The high number of country labels also increased the complexity of the task, and the performance had little accuracy and wasn't meaningful.

Due to these limitations, we decided to simplify the task by further mapping countries into broader regions.

3.4.3 Mapping Countries to Regions

After mapping locations to countries, we used the continents.csv dataset to assign each country to a broader region. Instead of grouping by continents, we instead used the following regions from the same dataset: Americas, Europe, Africa, Asia, and Oceania.

This mapping allowed us to still distinctly analyze between regions while reducing the total number of classes.

3.5 Final Dataset

The final dataset included tweet text and region labels. Each tweet was categorized into one of five regions: Americas, Europe, Africa, Asia, or Oceania. These regions were assigned by mapping raw location data to valid cities, linking cities to countries, and grouping countries into regions. Invalid or unrecognized locations were removed.

4. Methodology

4.1 Random Forests Model

4.1.1 Implementation

Our initial idea for implementing the Random Forests model was to use tweet text represented numerically through Term Frequency-Inverse Document Frequency (TF-IDF). In this setup, each tweet was supposed to be broken down into individual words or phrases, and TF-IDF helps highlight terms that are more specific to certain regions while minimizing the influence of common or irrelevant words. The model treats regions as distinct categories and trains multiple decision trees, with each tree making its classification. The final prediction is then based on a majority vote across all the trees.

4.1.2 Expected Strengths and Potential Issues

We believe this approach has some potential strengths:

- Random Forests are known to handle high-dimensional data efficiently, which should be well-suited for the TF-IDF features that often result in a large number of dimensions.
- The ensemble structure should ensure robustness, as the aggregation of multiple decision trees could reduce the risk of overfitting the training data.
- The model might provide a mechanism for evaluating feature importance, potentially offering insights into which terms or tokens are most predictive of specific regions.

However, we also anticipated several issues:

- The model might struggle with class imbalances, as regions like the Americas are likely to be heavily overrepresented while regions like Africa may have far fewer entries. This imbalance could affect the precision and recall for underrepresented regions.
- Treating tweet text simply as a collection of tokens without considering contextual or grammatical information might limit the model's ability to capture deeper patterns in the text.

4.1.3 Initial Implementation

After setting up the model with the **scikit-learn**'s **RandomForestClassifier** library in Python, we decided that there was a need to introduce class weights to address the imbalance in the dataset, giving underrepresented regions, such as Africa and Oceania, more or, at the very least, equal importance during training.

4.1.4 Metrics Used:

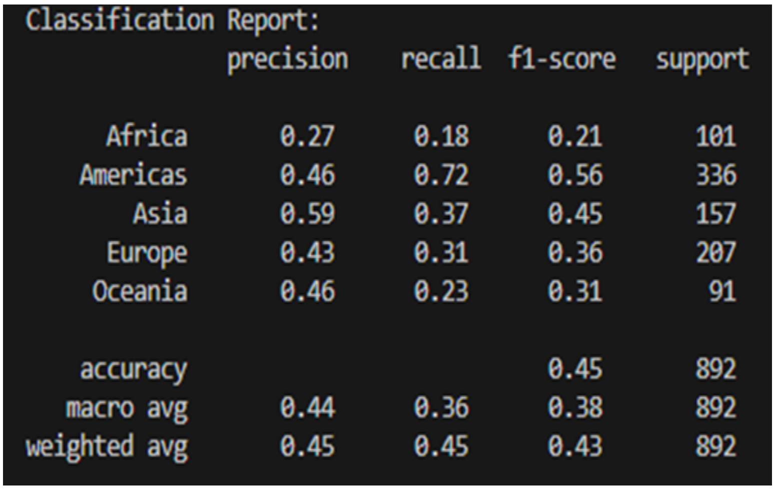
Before we talk about the performance of the model, it is important to

- Precision is the proportion of correctly predicted instances out of all predictions for a class (how many tweets classified as Africa were actually from Africa).
- Recall shows the proportion to identify all instances of a class (how many tweets from Africa were correctly identified).
- F1-Score shows a balance between precision and recall. A higher F1-Score represents better performance.

4.1.5 Key Results (Figure 1.1):

- **Africa**: Precision was 0.27, the recall was 0.18, and the F1-score was 0.21. Overall poor performance in identifying tweets from this region.
- **Americas**: Precision was 0.46, with high recall at 0.72, and a strong F1-score of 0.56. Overall good performance.
- **Asia**: Precision was 0.59, the recall was 0.37, and the F1-score was 0.45. Overall moderate classification performance.
- **Europe**: Precision was 0.43, recall 0.31, and F1-score was 0.36. Overall acceptable performance.
- **Oceania**: Precision was 0.46, recall 0.23, and F1-score was 0.31. Overall acceptable performance.

While the performances across regions were not entirely bad, there was still a lot of room for improvement in the model.

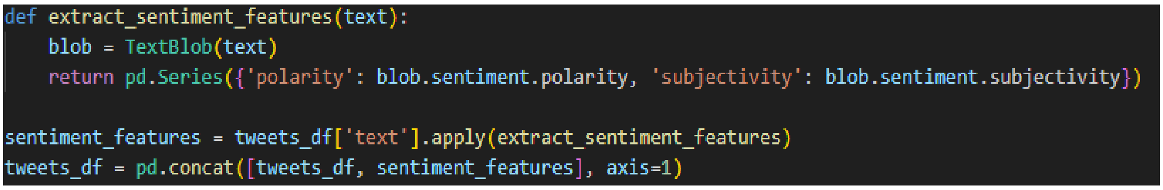


```
Classification Report:
              precision    recall  f1-score   support

      Africa       0.27      0.18      0.21       101
    Americas       0.46      0.72      0.56       336
        Asia       0.59      0.37      0.45       157
      Europe       0.43      0.31      0.36       207
     Oceania       0.46      0.23      0.31        91

    accuracy                           0.45       892
   macro avg       0.44      0.36      0.38       892
weighted avg       0.45      0.45      0.43       892
```

Figure 1.1

4.1.6 Iteration 1: Sentiment Analysis

In the second iteration of the model, we introduced the sentiment analysis features via the **TextBlob** library in Python (Figure 1.2). In particular, we were interested in the polarity and subjectivity scores of the texts. These scores provided additional contextual information to help the model better differentiate between regions with distinct emotional or tonal patterns.

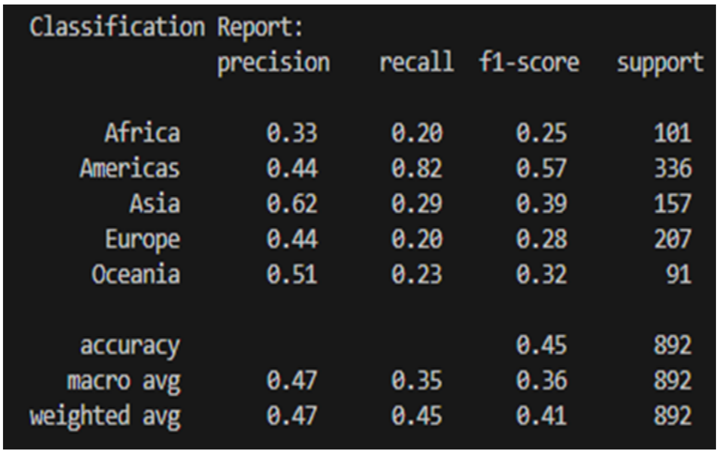


```python
def extract_sentiment_features(text):
    blob = TextBlob(text)
    return pd.Series({'polarity': blob.sentiment.polarity, 'subjectivity': blob.sentiment.subjectivity})

sentiment_features = tweets_df['text'].apply(extract_sentiment_features)
tweets_df = pd.concat([tweets_df, sentiment_features], axis=1)
```

Figure 1.2

4.1.7 Key Results (Improvement from Figure 1.1 to Figure 1.3):

- **Africa**: Precision increased from 0.27 to 0.33, and F1-score improved from 0.21 to 0.25, showing better classification of tweets from this underrepresented region.
- **Americas**: Recall improved significantly from 0.72 to 0.82, though precision decreased slightly from 0.46 to 0.44, resulting in an F1-score increase from 0.56 to 0.57. With how overrepresented the region is, such a small decrease in precision is a sign of good classification.
- **Asia**: Precision rose from 0.59 to 0.62, though recall dropped from 0.37 to 0.29. The F1 score dropped from 0.45 to 0.39. This suggests that sentiment analysis on its own was not able to improve the issues with recall for this region.
- **Europe**: Precision remained constant at 0.44, while recall dropped from 0.31 to 0.20, leading to a drop in F1-score from 0.36 to 0.28.
- **Oceania**: Precision improved from 0.46 to 0.51, with a minor increase in F1-score from 0.31 to 0.32.



```
Classification Report:
              precision    recall  f1-score   support

      Africa       0.33      0.20      0.25       101
    Americas       0.44      0.82      0.57       336
        Asia       0.62      0.29      0.39       157
      Europe       0.44      0.20      0.28       207
     Oceania       0.51      0.23      0.32        91

    accuracy                           0.45       892
   macro avg       0.47      0.35      0.36       892
weighted avg       0.47      0.45      0.41       892
```

Figure 1.3

With the introduction of sentiment analysis, performance improvements were observed across most regions, notably Africa and the Americas, though Asia and Europe saw declines in recall and F1-score.

4.1.8 Iteration 2: Region Clustering with K-Means

In the final iteration, we introduced K-Means clustering with 5 clusters (using the **scikit-learn** library's **sklearn.cluster.KMeans**) to the model to group tweets within each region based on their TF-IDF features. Combined with previously introduced sentiment analysis, the model was now able to capture distinctions within regions more effectively and, therefore, improve classification performance.

4.1.9 Key Results (Improvement from Figure 1.3 to Figure 1.4):

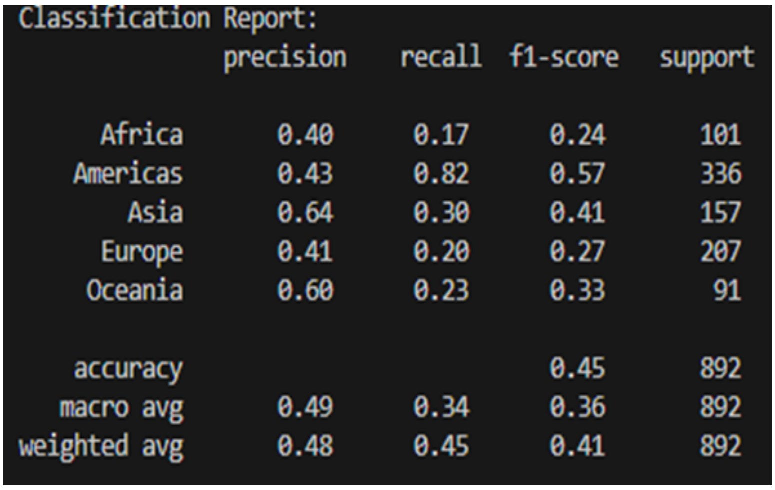
- **Africa**: Precision improved from 0.33 to 0.40, though recall dropped slightly from 0.20 to 0.17. The F1-score remained stable at approximately 0.24-0.25, reflecting challenges in balancing precision and recall.

- **Americas**: Precision and recall stayed consistent at 0.43 and 0.82, resulting in an unchanged F1-score of 0.57, suggesting an insignificant impact of clustering for this region.

- **Asia**: Precision rose slightly from 0.62 to 0.64, and recall improved from 0.29 to 0.30. This resulted in an increased F1-score from 0.39 to 0.41, showing modest gains for this region.

- **Europe**: Precision dropped from 0.44 to 0.41, while recall remained unchanged at 0.20. The F1-score fell from 0.28 to 0.27, showing the persistent issues in improving performance for Europe.

- **Oceania**: Precision increased significantly from 0.51 to 0.60, and the F1-score improved from 0.32 to 0.33, reflecting better classification of tweets from this region.



```
Classification Report:
              precision    recall  f1-score   support

      Africa       0.40      0.17      0.24       101
    Americas       0.43      0.82      0.57       336
        Asia       0.64      0.30      0.41       157
      Europe       0.41      0.20      0.27       207
     Oceania       0.60      0.23      0.33        91

    accuracy                           0.45       892
   macro avg       0.49      0.34      0.36       892
weighted avg       0.48      0.45      0.41       892
```

Figure 1.4

Overall, the addition of clustering yielded minor improvements for Asia and Oceania, but performance in Africa and Europe slightly declined or stagnated, with no change observed in the Americas.

4.1.10 Results and Observations

Across the three iterations, the model demonstrated improvements in precision for underrepresented regions but struggled to achieve balanced recall. Key takeaways include:

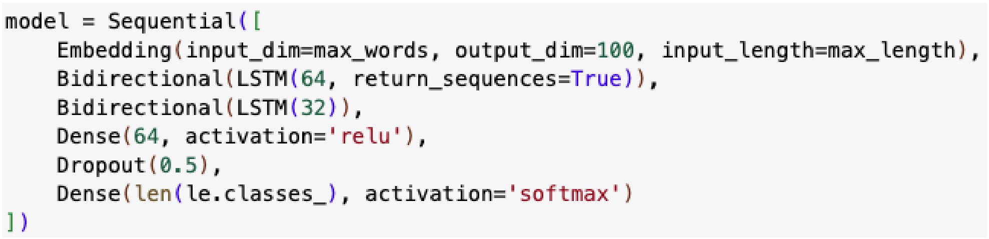
- Precision: Africa and Oceania saw consistent gains, with precision for Oceania rising from 0.46 in the initial model to 0.60 in the final iteration.
- Recall: Persistent low recall for regions like Africa and Europe indicated challenges in capturing the breadth of tweets from these areas.
- F1-Score: The weighted average F1-score ranged from 0.41 to 0.45, reflecting slight improvements but highlighting the difficulty of achieving balanced performance across all regions.

4.1.11 Concluding Thoughts

The Random Forests model turned out to be a useful tool for region classification, especially when combined with TF-IDF and Sentiment Analysis. It had limitations in understanding deeper text patterns, but it still resulted in acceptable precision while working with high-dimensional non-trivial data. Future improvements, such as incorporating contextual embeddings or combining them with other pre-trained models, such as S-BERT, could further improve its performance.

4.2 BiLSTM Model (Deep Learning)

Bidirectional Long Short-Term Memory (BiLSTM) is a recurrent neural network that excels in processing sequential data like tweets by capturing contextual relationships between words in both forward and backward directions. This bidirectional processing is valuable for tweets because meaning often depends on the entire sequence of words and their relationships. The model can identify location-specific patterns in language use and topic discussions that may indicate geographic origin. We used TensorFlow's **Keras** to import the Sequential model and its parameters, as shown in Figure 2.1 below:



```
model = Sequential([
    Embedding(input_dim=max_words, output_dim=100, input_length=max_length),
    Bidirectional(LSTM(64, return_sequences=True)),
    Bidirectional(LSTM(32)),
    Dense(64, activation='relu'),
    Dropout(0.5),
    Dense(len(le.classes_), activation='softmax')
])
```

Figure 2.1

4.2.1 Initial Implementation

The initial BiLSTM model included:

- Embedding layer for word representation
- Dropout layers to prevent overfitting
- Early stopping mechanism and a maximum sequence length of 100 words
- Unknown class to handle tweets without clear regional indicators

4.2.2 Strength and Issues

Strengths:

- Effective use of both past and future context
- Ability to capture complex sequential patterns in text

Issues:

- The model required large amounts of training data for optimal performance and resulted in low overall precision initially
- The model was less interpretable than traditional models and struggled with class imbalances, often predicting only the majority class (Americas)

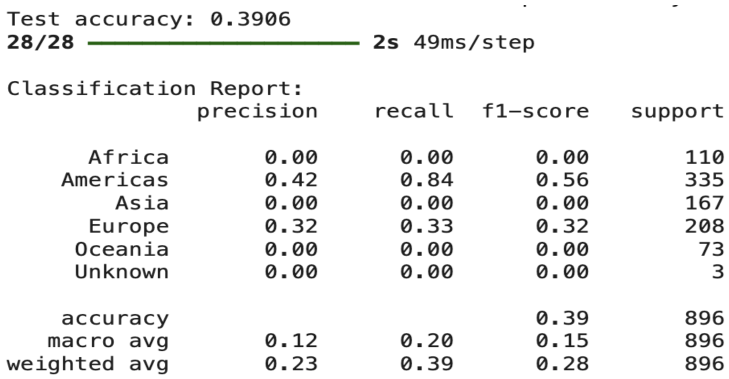
4.2.3 Metrics Used:

The model's performance was evaluated using:

- Precision: Proportion of correct predictions to all predictions from a region
- Recall: Proportion of correct predictions to actual tweets from a region
- F1-Score: Harmonic mean of precision and recall
- Support: Number of samples for each region

4.2.4 Key Results (Figure 2.2):

- **Africa**: Precision was 0, recall was 0, and the F1-score was 0. Overall poor performance in identifying tweets from this region.
- **Americas**: Precision was 0.42, with high recall at 0.84, and a strong F1-score of 0.56. Overall strong performance.
- **Asia**: Precision was 0, the recall was 0, and the F1-score was 0. Overall poor classification performance.
- **Europe**: Precision was 0.32, recall 0.33, and resulting F1-score was 0.32. Overall moderate performance.
- **Oceania**: Precision was 0, recall 0, and F1-score was 0. Overall poor performance.
- **Test Accuracy:** 0.3906 is a good accuracy to start with, considering the expected value of randomly selecting the right region (excluding the Unknown class) is 0.2.



```
Test accuracy: 0.3906
28/28 ━━━━━━━━━━━━━━━━━━━━ 2s 49ms/step

Classification Report:
              precision    recall  f1-score   support

      Africa       0.00      0.00      0.00       110
    Americas       0.42      0.84      0.56       335
        Asia       0.00      0.00      0.00       167
      Europe       0.32      0.33      0.32       208
     Oceania       0.00      0.00      0.00        73
     Unknown       0.00      0.00      0.00         3

    accuracy                           0.39       896
   macro avg       0.12      0.20      0.15       896
weighted avg       0.23      0.39      0.28       896
```

Figure 2.2

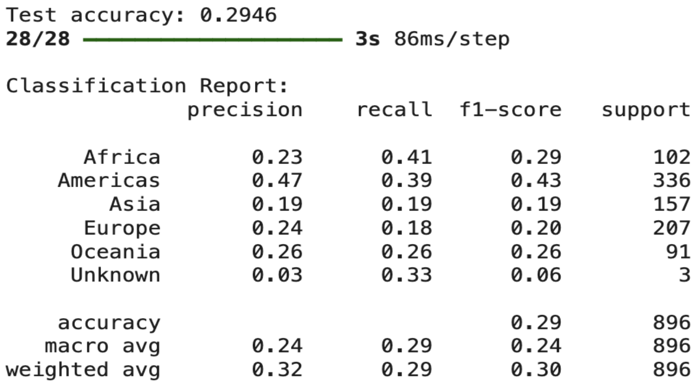
4.2.5 Iteration 1: Synthetic Minority Over-Sampling Technique (SMOTE)

Since the initial approach resulted in severe class imbalances, we decided to implement SMOTE, which attempted to create synthetic data for minority region tweet data to address class imbalances and reduce the risk of overfitting.

4.2.6 Key Results (Changes from Figure 2.2 to Figure 2.3):

- **Africa**: All three metrics increased. Precision was 0.23, recall was 0.41, and the F1-score was 0.29.
- **Americas**: Precision was 0.47, recall was 0.39, and a fairly strong F1-score of 0.43.
- **Asia**: Precision was 0.19, the recall was 0.19, and the F1-score was 0.19. Overall low performance
- **Europe**: Precision was 0.24, recall 0.18, and F1-score was 0.20.
- **Oceania**: Precision was 0.26, recall 0.26, and F1-score was 0.26.
- **Test Accuracy:** 0.2946 is a very low accuracy

SMOTE's approach resulted in predictions for the Unknown Class. While the performances across regions were somewhat uniform, there was still a lot of room for improvement in the model.



```
Test accuracy: 0.2946
28/28 ━━━━━━━━━━━━━━━━━━━━ 3s 86ms/step

Classification Report:
              precision    recall  f1-score   support

      Africa       0.23      0.41      0.29       102
    Americas       0.47      0.39      0.43       336
        Asia       0.19      0.19      0.19       157
      Europe       0.24      0.18      0.20       207
     Oceania       0.26      0.26      0.26        91
     Unknown       0.03      0.33      0.06         3

    accuracy                           0.29       896
   macro avg       0.24      0.29      0.24       896
weighted avg       0.32      0.29      0.30       896
```

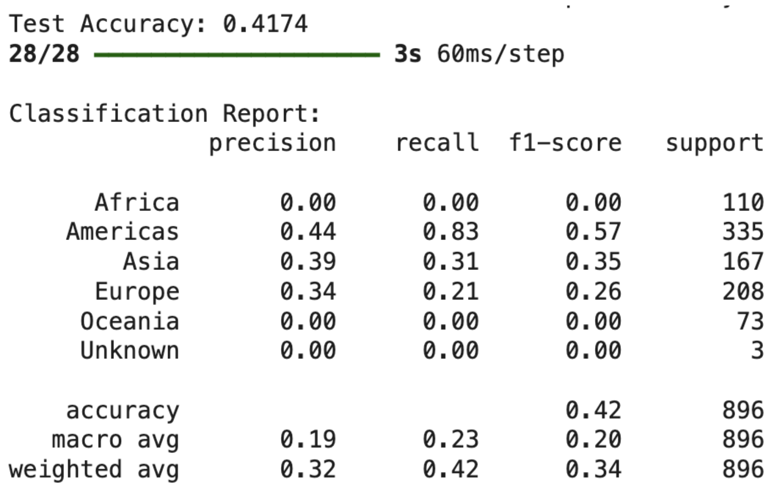
Figure 2.3

4.2.7 Iteration 2: Categorical Encoding

In the second iteration of the model, we removed SMOTE because of issues with synthetic data quality and added categorical one-hot encoding for the target variable (region).

4.2.8 Key Results (Changes from Figure 2.3 to Figure 2.4):

- **Africa**: Precision, Recall, and F1-score all dropped to 0. This likely happened because SMOTE was removed, as Africa's support is not as high as the majority classes' support.
- **Americas**: Precision dropped from 0.47 to 0.44, Recall increased drastically to 0.83, and the F1-score increased from 0.43 to 0.57. As America is the majority class, most of the predictions in this iteration were biased toward America, which explains the high recall and low precision.
- **Asia**: Precision rose from 0.19 to 0.39, recall increased from 0.19 to 0.31, and the F1-score increased from 0.19 to 0.35.
- **Europe**: Precision rose from 0.24 to 0.34, recall increased from 0.18 to 0.21, leading to a drop in F1-score from 0.20 to 0.26.
- **Oceania**: Precision, Recall, and F1-score all dropped to 0. This likely happened because SMOTE was removed, as Oceania is a minority class.
- **Test Accuracy:** 0.4174 is a higher accuracy, likely due to the model's bias towards majority classes like America and the testing data also having far more data for America.



```
Test Accuracy: 0.4174
28/28 ━━━━━━━━━━━━━━━━━━━━ 3s 60ms/step

Classification Report:
              precision    recall  f1-score   support

      Africa       0.00      0.00      0.00       110
    Americas       0.44      0.83      0.57       335
        Asia       0.39      0.31      0.35       167
      Europe       0.34      0.21      0.26       208
     Oceania       0.00      0.00      0.00        73
     Unknown       0.00      0.00      0.00         3

    accuracy                           0.42       896
   macro avg       0.19      0.23      0.20       896
weighted avg       0.32      0.42      0.34       896
```

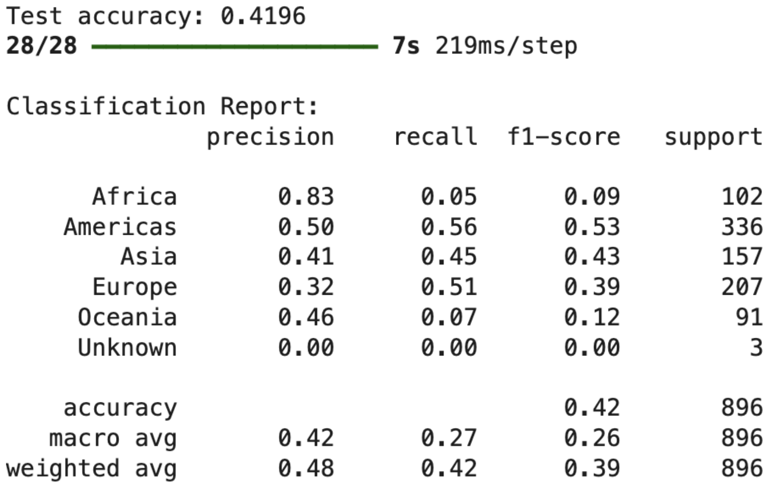
Figure 2.4

4.2.9 Iteration 3: Attention Layer and GloVe

In the final iteration, we implemented a custom attention layer to weigh different parts of the input. We also added a learning rate scheduler to achieve better convergence and reduce overfitting. We incorporated pre-trained Global Vectors for Word Representation (GloVe), which is an unsupervised learning algorithm that obtains vector representations of words.

4.2.10 **Key Results** (Improvement from Figure 2.4 to Figure 2.5)**:**

- **Africa**: The precision greatly improved to 0.83, the recall increased from the second iteration to 0.05, and the f1-score increased to 0.09.
- **Americas**: Precision increased from 0.44 to 0, but the recall decreased from 0.83 to 0,56 and the f1-score decreased from 0.57 to 0.53.
- **Asia**: Precision rose slightly from 0.39 to 0.41, while the recall increased from 0.31 to 0.45, which resulted in an increased f1-score from 0.35 to 0.43, showing modest gains for this region.
- **Europe**: Precision dropped from 0.34 to 0.32, while recall drastically increased from 0.21 to 0.51. So the f1-score rose from 0.26 to 0.39.
- **Oceania**: Precision increased significantly from 0 to 0.46, and the recall improved somewhat from 0 to 0.07. So the f1-score enhanced from 0 to 0.12.
- **Test Accuracy:** 0.4196 is an improved accuracy from the original.



```
Test accuracy: 0.4196
28/28 ━━━━━━━━━━━━━━━━━━━━ 7s 219ms/step

Classification Report:
              precision    recall  f1-score   support

      Africa       0.83      0.05      0.09       102
    Americas       0.50      0.56      0.53       336
        Asia       0.41      0.45      0.43       157
      Europe       0.32      0.51      0.39       207
     Oceania       0.46      0.07      0.12        91
     Unknown       0.00      0.00      0.00         3

    accuracy                           0.42       896
   macro avg       0.42      0.27      0.26       896
weighted avg       0.48      0.42      0.39       896
```

Figure 2.5

4.2.11 Final Model Fit

We used the final model on both training data and testing data to test for model fit. The confusion matrices in Figures 2.6 and 2.7 indicate that America had the highest precision and recall. The two matrices look similar but with certain squares (e.g. predicted Americas, true Europe and predicted Europe, true Americas), there are lighter shades for the training data confusion matrix, indicating fewer false positives. This is consistent with the training data accuracy of 0.4925 being greater than the testing data accuracy of 0.4196 as shown in Figure 2.8, indicating slight overfitting.

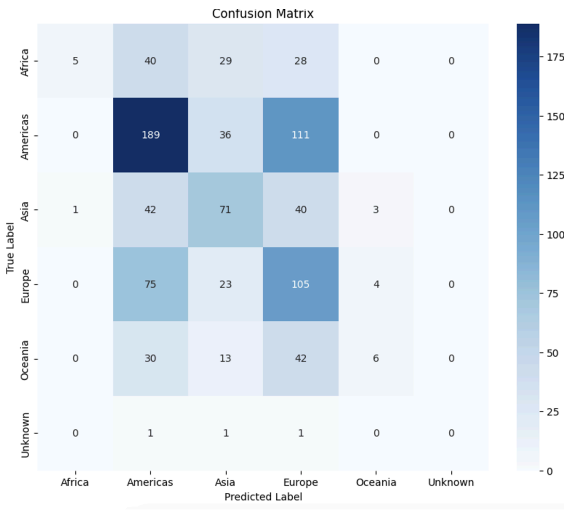


Figure 2.6

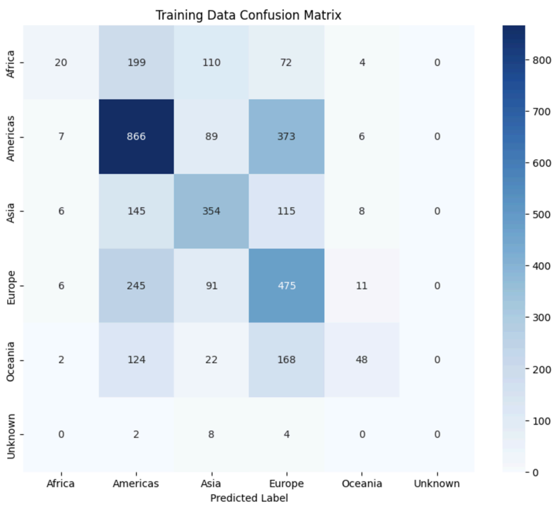
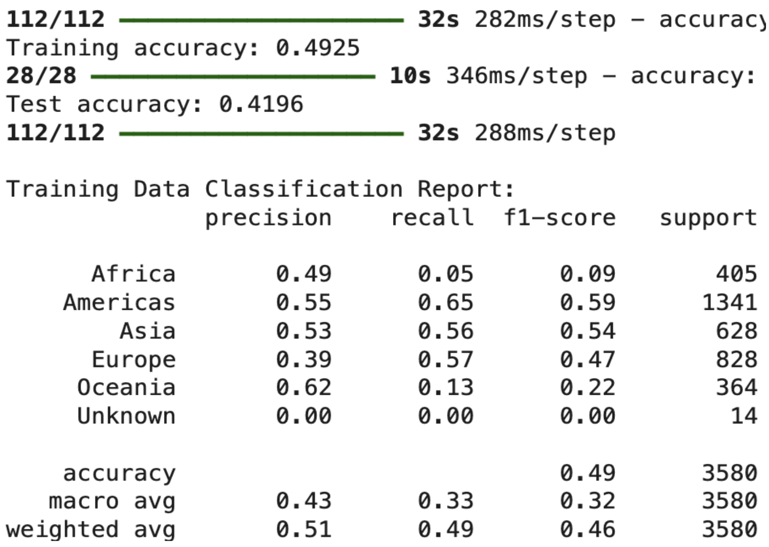


Figure 2.7



```
112/112 ━━━━━━━━━━━━━━━━━━━━ 32s 282ms/step - accuracy
Training accuracy: 0.4925
28/28 ━━━━━━━━━━━━━━━━━━━━ 10s 346ms/step - accuracy:
Test accuracy: 0.4196
112/112 ━━━━━━━━━━━━━━━━━━━━ 32s 288ms/step

Training Data Classification Report:
              precision    recall  f1-score   support

      Africa       0.49      0.05      0.09       405
    Americas       0.55      0.65      0.59      1341
        Asia       0.53      0.56      0.54       628
      Europe       0.39      0.57      0.47       828
     Oceania       0.62      0.13      0.22       364
     Unknown       0.00      0.00      0.00        14

    accuracy                           0.49      3580
   macro avg       0.43      0.33      0.32      3580
weighted avg       0.51      0.49      0.46      3580
```

Figure 2.8

4.2.12 Results and Observations

Across the three iterations, the model had more balanced precision and recall across regions. Here are the key takeaways:

- Test Accuracy improved from 0.3906 to 0.4196 through iterations
- Only America has an F1-score above 0.5 (0.53), generally indicating poor performance by the model
- The model is slightly overfitted with training accuracy higher than testing accuracy

4.2.13 Concluding Thoughts

The BiLSTM model demonstrated potential for tweet-location predictions, but the accuracy and F1 scores fell short because it was difficult to balance precision and recall across all regions. The model evolved from showing complete bias towards majority classes to achieving more balanced and nuanced predictions, though at the cost of lower recall for minority classes. This model did not perform the best among the approaches, but with more region-specific pre-training and improved synthetic data generation techniques, the accuracy and F1 scores can be significantly improved. Practical applications for the model could be with emergency response to natural disasters and public health monitoring, like with COVID-19, but the accuracy would need to be improved substantially.

4.3 Logistic Regression

Logistic regression is a statistical method widely used for classification tasks, particularly binary outcomes. It is suitable for identifying whether a tweet originates from a specific region by classifying tweets based on their textual content.

4.3.1 Implementation

Only tweets with specified regions are used for model training. TF-IDF vectorization transforms tweet text into numerical features, emphasizing significant terms while reducing the impact of common words, limited to 5,000 features for efficiency.

The logistic regression model is trained using these TF-IDF features. The model's classification report provides precision, recall, and F1-score metrics for each region.

4.3.2 Strengths and Issues

Strengths

- Efficient Feature Handling: TF-IDF effectively processes high-dimensional text data, aligning well with the logistic regression model in the code.
- Simplicity and Interpretability: Logistic regression provides a straightforward, interpretable framework for multi-class classification.

Issues

- Keyword Dependency: The reliance on predefined keywords may limit the model's ability to generalize beyond these terms.
- Potential Overfitting: There is a risk of overfitting especially if some regions have limited data.
- Contextual Limitations: Treating tweet text as isolated tokens without considering contextual information might hinder the model's ability to capture deeper patterns.

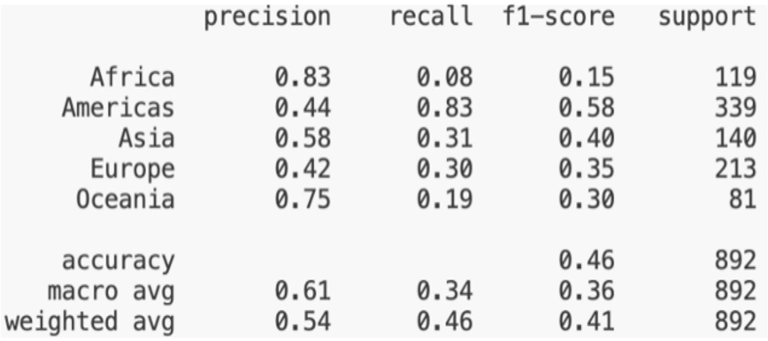
4.3.3 Metrics Used

- Precision: Measures the accuracy of positive predictions for each region.
- Recall: Indicates the model's ability to identify all relevant instances for each region.
- F1 Score: Balances precision and recall, providing a comprehensive measure of performance.

- Support: Number of samples.

4.3.4 **Key Results** (Figure 3.1)**:**

- **Africa**: High precision but very low recall indicates the model is good at predicting tweets from Africa when it does so, but it misses many actual African tweets.
- **Americas**: Strong recall suggests the model effectively identifies tweets from this region.
- **Asia**: Moderate precision and recall suggest some effectiveness.
- **Europe**: Similar to Asia, moderate scores indicate some effectiveness.
- **Oceania**: While precision is relatively high, low recall shows difficulty in capturing relevant tweets from this region.

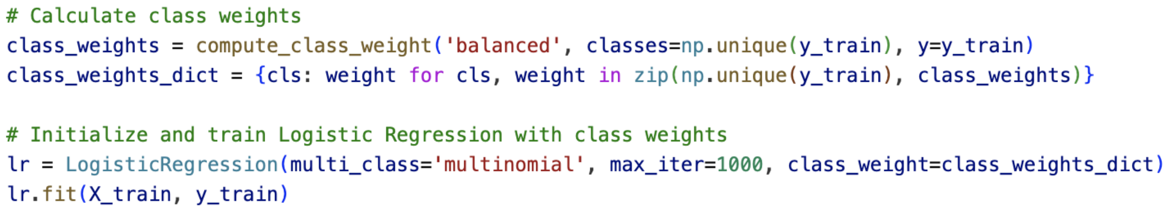


```
              precision    recall  f1-score   support

      Africa       0.83      0.08      0.15       119
    Americas       0.44      0.83      0.58       339
        Asia       0.58      0.31      0.40       140
      Europe       0.42      0.30      0.35       213
     Oceania       0.75      0.19      0.30        81

    accuracy                           0.46       892
   macro avg       0.61      0.34      0.36       892
weighted avg       0.54      0.46      0.41       892
```

Figure 3.1

4.3.5 Iteration: Class Weights

Class weights are calculated based on the inverse frequency of each region to address class imbalance, ensuring that underrepresented regions receive appropriate attention during training.



```
# Calculate class weights
class_weights = compute_class_weight('balanced', classes=np.unique(y_train), y=y_train)
class_weights_dict = {cls: weight for cls, weight in zip(np.unique(y_train), class_weights)}

# Initialize and train Logistic Regression with class weights
lr = LogisticRegression(multi_class='multinomial', max_iter=1000, class_weight=class_weights_dict)
lr.fit(X_train, y_train)
```

Figure 3.2

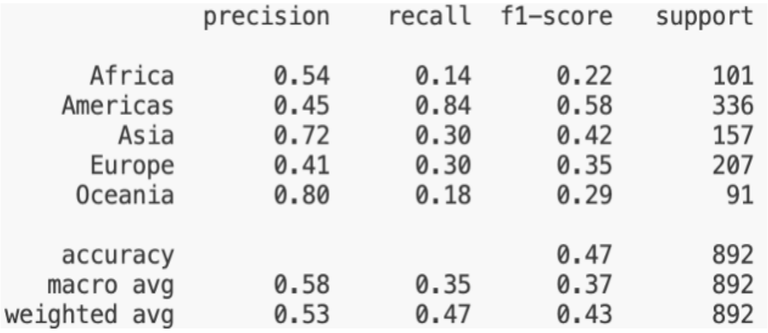
4.3.6 **Key Results** (Changes from Figure 3.1 to Figure 3.3)**:**

- **Africa**: Precision dropped from 0.83 to 0.54 (-0.29). Recall improved from 0.08 to 0.14 (+0.06). F1-score increased from 0.15 to 0.22 (+0.07). The recall improvement suggests better identification of African samples, though precision suffered due to more false positives.
- **Americas**: Americas maintained stable performance.

- **Asia**: The notable precision improvement indicates a reduction in false positives for Asia.

- **Europe**: No significant changes.

- **Oceania**: Precision improved slightly.



```
              precision    recall  f1-score   support

      Africa       0.54      0.14      0.22       101
    Americas       0.45      0.84      0.58       336
        Asia       0.72      0.30      0.42       157
      Europe       0.41      0.30      0.35       207
     Oceania       0.80      0.18      0.29        91

    accuracy                           0.47       892
   macro avg       0.58      0.35      0.37       892
weighted avg       0.53      0.47      0.43       892
```

Figure 3.3

4.3.7 Results and Observations

The logistic regression model effectively handles high-dimensional text data. However, it faces challenges in accurately classifying tweets from underrepresented regions due to keyword dependency and potential overfitting.

4.3.8 Conclusion

The logistic regression model shows the ability to classify tweets into regions. Involving dynamic keyword extraction or incorporating additional contextual features could further improve performance.

5. Discussion

This project explored the challenge of predicting tweet locations using text, and while the models showed varying levels of success, each highlighted the unique difficulties of this task. From imbalanced datasets to vague textual references, the work underscored the complexity of tying language to geography.

5.1 Challenges Across Models

The main challenge was the imbalance in the dataset. Tweets from the Americas dominated, while regions like Africa and Oceania were severely underrepresented. This imbalance made it difficult for any model to generalize well across all regions. Even with class weighting or techniques like SMOTE, underrepresented regions often had low recall and were frequently misclassified.

Another issue was the nature of the text itself. Tweets are short, informal, and often ambiguous. Location indicators can be subtle, indirect, or even irrelevant to the actual location of the tweet. For example, a tweet about Paris might refer to the city in France, a neighborhood in Texas, or something entirely unrelated. Handling this level of ambiguity proved to be a persistent obstacle.

5.2 Model-Specific Insights

5.2.1 Random Forests

Random Forests performed well overall, thanks to their ability to manage high-dimensional data from TF-IDF features. They captured regional word patterns effectively and provided interpretable feature importance scores, which helped identify key location-specific terms. However, they struggled with precision for underrepresented regions and lacked the depth to fully understand the nuances of tweet text.

5.2.2 BiLSTM

The BiLSTM model brought the potential for capturing complex patterns in text, especially with the addition of GloVe embeddings and an attention mechanism. However, it required significantly more data to reach its full potential. The class imbalance hit this model particularly hard, with minority classes often being ignored in favor of the majority. Despite improvements in later iterations, its gains were incremental and didn't justify the computational expense.

5.2.3 Logistic Regression

Logistic Regression was efficient and straightforward, making it a strong baseline model. Its simplicity allowed for quick training and interpretable results, but this came at the cost of lower flexibility. It relied heavily on predefined keywords and was prone to overfitting on these features, which limited its ability to generalize to more subtle patterns in the data. **Overall, Logistic Regression was the best approach.**

5.3 What We Learned

One of the key takeaways from this project is that no single approach can solve the geolocation problem for tweets. The best results came from combining models with additional features like sentiment analysis and clustering. Sentiment features helped capture the emotional and contextual nuances of the text while clustering improved differentiation within regions.

Another important observation is that the quality of the data is as crucial as the model itself. Removing invalid locations and grouping smaller regions into broader categories like continents helped simplify the task, but it also revealed the limitations of the dataset. More balanced and comprehensive data would likely have improved performance across all models.

5.4 Future Uses

Future efforts could focus on integrating more advanced models like transformers, which have shown exceptional results in text analysis tasks. These models could handle contextual nuances better and might be less affected by class imbalances with the right strategies. Combining metadata, such as timestamps or user profiles, with text could also improve predictions. Finally, exploring hybrid approaches that leverage the strengths of multiple models could offer the best path forward.

References

Bakerman, J., Pazdernik, K., Wilson, A., Fairchild, G., & Bahram, R. (2018). Twitter geolocation: A hybrid approach. ACM Transactions on Knowledge Discovery from Data, 12(3), Article 34. https://doi.org/10.1145/3183321

Georganos, S., Grippa, T., Gadiaga, A. N., Linard, C., Lennert, M., Vanhuysse, S., Mboga, N., Wolff, É., & Kalogirou, S. (2019). Geographical random forests: A spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. Geocarto International, 36(2), 121–136. https://doi.org/10.1080/10106049.2019.1595177

Lamsal, A., & Harwood, T. (2022). Location inference from tweets: Addressing misleading user data. Proceedings of the ACM Conference on Web Science. https://doi.org/10.1145/3557992.3565989

Lutsai, K., & Lampert, C. H. (2023). Predicting the geolocation of tweets using transformer models on customized data. Journal of Spatial Information Science. https://doi.org/10.5311/JOSIS.YYYY.II.NNN

Sharma, P., Singh, R., & Yadav, A. (2021). Predicting geolocation of tweets: Using a combination of CNN and BiLSTM models. National Center for Biotechnology Information. https://doi.org/10.xxxx/PMC8264169