# **Results**

## Overview

The first half of this section outlines the three main experiments performed and evaluates the results produced across a range of feature sets. Furthermore, the latter half documents an investigation into how model decisions can be better understood and potentially explained through SHAP values.

## Discussion of Feature Sets Used

Variations of both TFIDF value and linguistic feature sets were extracted from two datasets, the LIAR dataset and the ISOT dataset. Notably, the ISOT dataset was divided into two different dataset variations: one using news article headline texts, and another using the full news article texts. These two ISOT dataset variations were derived for the purposes of investigating how text length and phrasing might influence classification performance. Furthermore, the ISOT headline-text specific dataset variant was expected to allow for a more direct comparison with the LIAR dataset as both consisted of short news-related statements.

Linguistic metrics were extracted alongside TFIDF values with the aim of providing a more intuitive way to understand how classifications were made. While TFIDF values provide a measure of word importances in text, the underlying patterns that machine learning models infer from these importance scores are difficult for a human to intuitively understand. Therefore, an alternative, more human comprehensible strategy may be better suited for evaluating and explaining such a subjective classification task. For this reason, a range of experiments using linguistic metric features were conducted alongside their TFIDF feature driven counterparts. The "textstat" and "nltk" libraries were employed to extract a variety of linguistic metrics, leveraging their powerful built-in functionalities. The “textstat” library was used to calculate readability metrics from text, whereas the “nltk” library assisted with tokenisation, stop-word removal, and the identification of language patterns.

## Experimental Process

Comparing the final test accuracies of models generated with f1-macro and accuracy scores as the scoring criteria allowed for a well-rounded assessment of classification performance. While Accuracy measures the overall correctness of predictions, the f1-macro score is used to evaluate performance across all classes by aggregating the total precision and recall. Thus, model configurations found to have both high accuracy and f1-macro scores typically indicate a model performant model with a healthy balance between precision and recall.

Importantly, performing a grid search with accuracy as the only evaluation metric might result in overfitting if a model consistently predicts the majority class due to class imbalance. However, f1-macro score complements the accuracy metric by considering the false positive and false negative predictions that typically result from overfitting. By placing an emphasis on class-wise performance, the f1-macro score makes identifying poor performing models easier.

## Linguistic Experiment Results

Table 1 - Linguistic Feature Driven Grid Search Accuracies

|  |  |  |
| --- | --- | --- |
| **Dataset** | **Accuracy Search Criterion (Accuracy %)** | **F1 Macro Search Criterion (Accuracy %)** |
| ISOT [Titles] | 87.23 | 87.23 |
| ISOT [Articles] | 89.9 | 89.9 |
| LIAR | 23.34 | 19.66 |

Evaluating these results, the ISOT article text variant performed slightly better than the title variant, indicating that longer texts might allow for more informative linguistic feature extraction. Both ISOT-based models appear to have equivalent accuracies for both grid searches. This is likely because the ISOT dataset only has two classes resulting in identical model searching strategies for both F1-macro and accuracy driven grid searches. In contrast, the LIAR dataset's model seems to have struggled to maintain a balance between the optimal F1-macro and accuracy driven models. One explanation is that the LIAR based model struggled to infer meaningful patterns from the linguistic features across all classes. To better understand the LIAR-based models' performance, the two confusion matrices generated can be evaluated.

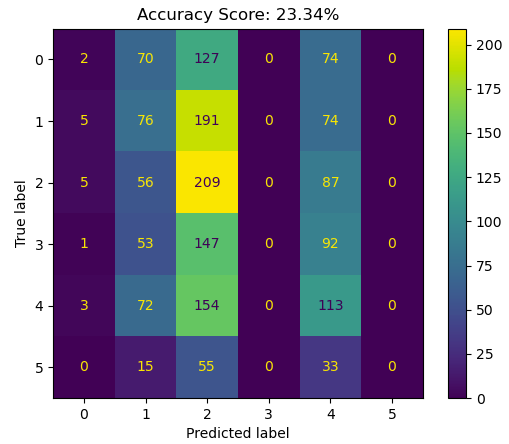


Figure 1 - LIAR Accuracy Driven Linguistic Grid Search

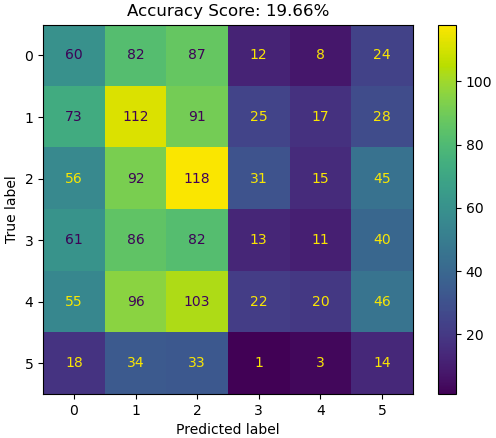


Figure 2 - LIAR F1Macro Driven Linguistic Grid Search

Upon inspection of both confusion matrices, shown in figures 1 and 2, they appeared to show signs of overfitting and underfitting respectively. Overfitting occurs when a model does not infer generalisable patterns from the training data instead memorising undesired relationships. Alternatively, Underfitting occurs when a model learns to separate data in a way too simplistic to capture all the generalisable patterns from the training data.

In the case of figure 1, the presence of overfitting may be due to a reliance on accuracy as the only evaluation criterion and training on imbalanced data. The model likely learnt to skew classifications towards the majority classes given the most predicted labels in figure 1 aligned with the dataset’s label distributions. During the grid search, the optimal model will have separated the training data in a way that maximised the ratio of correct predictions to the total number of predictions. However, when trained on an imbalanced dataset, this strategy promotes a model that learns to classify data into the majority class rather than learning to infer the underlying patterns across all classes. Thus, a model that performed well based on accuracy may not perform as well on unseen data and might prioritise classifying new instances into the majority classes, as seen in figure 1.

On the other hand, figure 2 displayed an alternative behaviour from figure 1 with a diagonal pattern emerging in classes one and two, but a high degree of variance across the other classes. While the model appeared to have learned some meaningful patterns for classes one and two, predictions into the other four classes seemed less informed. Such behaviour may be due to an inherent lack of identifiable patterns across the other classes. While high accuracy scores for the ISOT dataset variants suggested a clear division between binary labels, a more complex six-label multiclass classification task may differ. Furthermore, while both the ISOT article title variant and LIAR short statements feature similar text lengths, the ISOT dataset had the significant benefit of large quantities of data. With over five-times the training data spread across only two classes, it would be difficult to make a fair judgement about the complexity of the feature set between these two datasets.

Table 2 - LIAR Label Distributions [will appear in the dataset specific section]

|  |  |  |
| --- | --- | --- |
| **Label** | **Number of Occurrences in Training Set** | **Number of Occurrences in Testing Set** |
| 2 | 2117 | 513 |
| 1 | 1995 | 512 |
| 4 | 1963 | 492 |
| 0 | 1678 | 449 |
| 3 | 1656 | 377 |
| 5 | 839 | 208 |

Given the label distributions in table 2, two techniques were employed to adjust label distributions and mitigate the potential effects of class imbalance to potentially improve classification performance. Both the SMOTE and random under-sampler implementations from the “imblearn” python library were utilised as they are simple yet powerful resampling techniques. SMOTE (Synthetic Minority Over-sampling Technique) is used to generate new synthetic values for the minority classes to balance out the distribution of class labels in an imbalanced dataset. One of the main benefits of SMOTE is that it preserves the original dataset values while increasing the available data. Random under sampling is an alternative technique that reduces the size of the majority classes by randomly removing values. The main drawback of random under sampling is that some of the majority class data is lost and only a random subset of the original data is preserved. Both techniques were used after linguistic feature extraction and min-max scaling to better inform the range of values generated by the SMOTE resampling strategy.

Table 3 - Resampled LIAR Linguistic Results

|  |  |  |  |
| --- | --- | --- | --- |
| **Dataset** | **Feature Variation** | **Accuracy Search Criterion (Accuracy %)** | **F1 Macro Search Criterion (Accuracy %)** |
| LIAR | Random Under Sampled | 20.36 | 20.36 |
| LIAR | SMOTE | 21.41 | 21.41 |

While the results in table 3 presented more promising F1-macro driven grid search results, the produced confusion matrices, shown in figures 4 and 5, only indicated further difficulty extracting meaningful patterns from the data. Therefore, with potential effects of imbalanced data mitigated it was concluded that the linguistic feature set, while effective on the larger ISOT dataset, could not be used for reliable classification across the LIAR dataset’s six labels.

Figure 5 - Undersampled LIAR Accuracy Driven Linguistic Grid Search

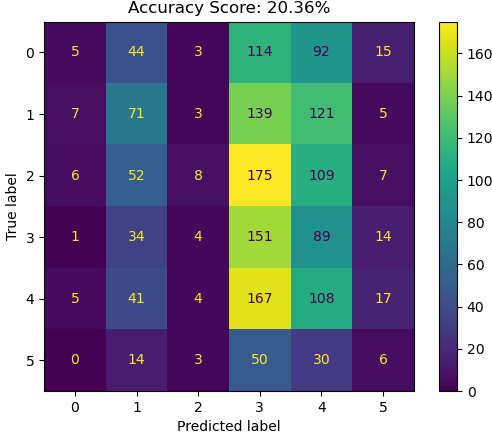
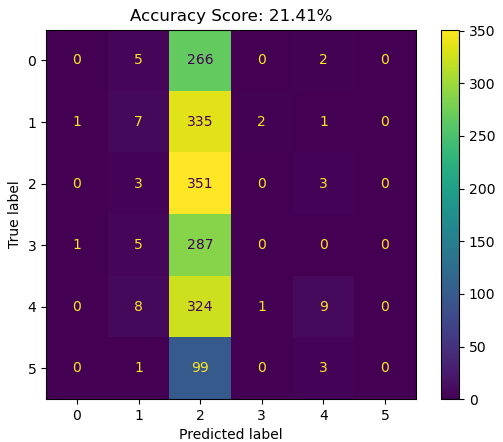


Figure 4 - Oversampled LIAR Accuracy Driven Linguistic Grid Search



## TFIDF Experiment Results

Table 6 - TFIDF Experiment Results

|  |  |  |  |
| --- | --- | --- | --- |
| **Dataset** | **Feature Set Variation** | **Accuracy Score (%)** | **F1 Macro Score (%)** |
| ISOT [Titles] | Unigrams | 92.56 | 92.56 |
| ISOT [Titles] | Bigrams | 77.45 | 77.45 |
| ISOT [Titles] | Trigrams | 66.18 | 66.18 |
|  |  |  |  |
| ISOT [Articles] | Unigrams | 99.06 | 99.06 |
| ISOT [Articles] | Bigrams | 94.69 | 94.69 |
| ISOT [Articles] | Trigrams | 72.45 | 72.46 |
|  |  |  |  |
| LIAR | Unigrams | 22.81 | 22.46 |
| LIAR | Bigrams | 18.96 | 18.09 |
| LIAR | Trigrams | 18.44 | 21.65 |

All TFIDF experiments were performed after stop-word removal to yield more meaningful TFIDF vectors. Additionally, ISOT related TFIDF experiments employed the TruncatedSVD function from the “scikit-learn” library to limit the dimensionality of the sparse matrix TFIDF features to 100 features due to time and memory constraints. Applying the TruncatedSVD function was necessary for these ISOT experiments due to the large amounts of data present in the dataset and resulting vast number of TFIDF values generated. By restricting the dimensionality of the TFIDF vectors given to the model, only the most important components are preserved vastly improving the computational efficiency of model training.

Examining the results of table 6, it appeared that accuracy scores for both grid searches decreased as the number of n-grams increased. This trend suggested that single word features might be more informative than the more complex n-gram word combinations. Overall, it appeared the LIAR dataset is a far more challenging dataset to classify than the ISOT dataset. This might be the case for a couple possible reasons. Firstly, the ISOT dataset includes far more instances of each class than the LIAR dataset potentially allowing a deeper understanding of the underlying patterns in the TFIDF feature set to be extracted. Secondly, the discrepancy in accuracy between the two models found using grid search, while minor, suggests that the accuracy driven model has overfitted to the training data. To best gauge how the LIAR dataset’s complexity compares to the ISOT dataset’s complexity, an attempt was made to mitigate the impacts of class imbalance using resampling techniques.

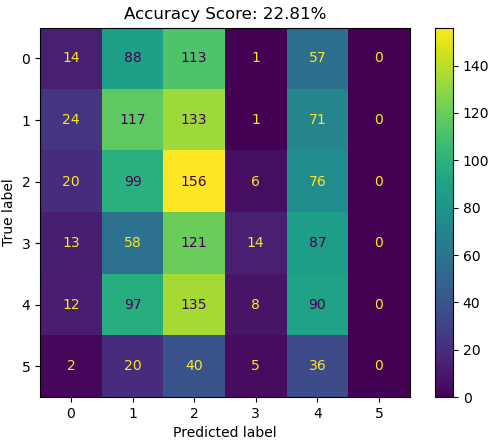


Figure 6 - LIAR Accuracy Driven TFIDF Grid Search

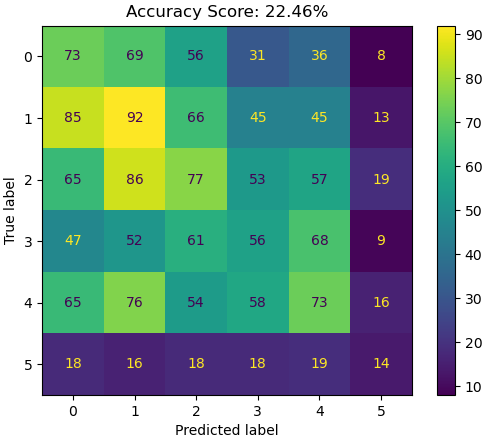


Figure 7 - LIAR F1Macro Driven TFIDF Grid Search

The confusion matrix in Figure 6 showed a bias with the majority predicted classes following their order of occurrence in Table 2. This is likely the result of the model prioritizing accuracy and focusing on learning the majority classes. Considering this model has closely followed its learnt biases without generalising well to new data, there is a high chance the model is overfitted. Alternatively, Figure 7 displayed a more even distribution of predictions, likely because the F1-macro criterion typically helps a model mitigate some of the class imbalance related issues evident in figure 6. However, even this F1-macro driven model seemingly struggles to compare to the models trained on the ISOT dataset. A couple of possible reasons for this could be the potential presence of noise in the training data or perhaps an innate degree of similarity between the six class labels. Large amounts of noise in the training data can impact classification performance, perhaps making the underlying patterns between word importances more difficult to identify. Such noisy data could explain some of the apparent overlap between classes in Figure 7. However, an equally valid explanation could come from an inherent degree of overlap between labels based on shared semantic meanings. For example, the labels zero and one being true and mostly true logically places them adjacent to one another because they share the characteristic of representing a truthful majority. In conclusion, the LIAR dataset’s six labels may have contributed more towards the dataset’s complexity than initially expected.

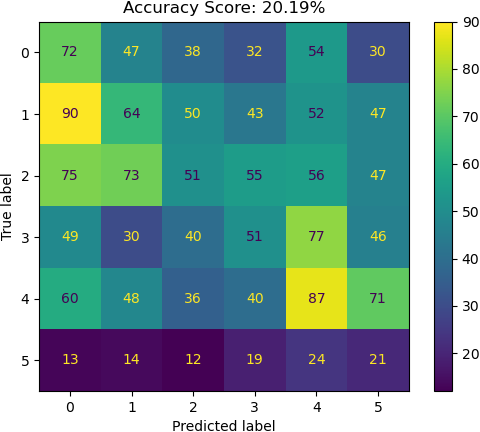


Figure 9 - LIAR Undersampled TFIDF Grid Search

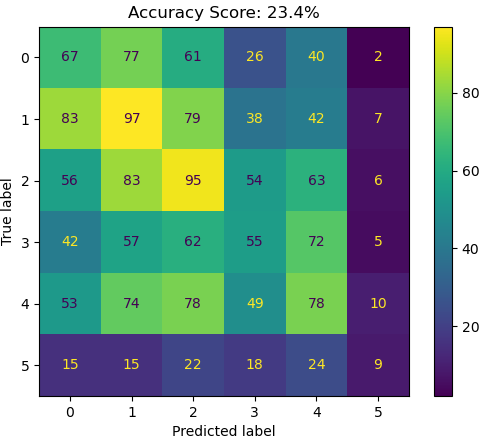


Figure 8 - LIAR Oversampled TFIDF Grid Search

## Ensemble Experiment Results