# **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. Additionally, 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 headlines, and another using the full news articles. The two ISOT dataset variations were derived for the purposes of investigating how text length and phrasing might influence classification performance. Furthermore, the ISOT headline 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 this data can be difficult for a human to understand. Therefore, an alternative, more human comprehensible strategy may be better suited for evaluating and explaining such a subjective classification task. Thus, a range of experiments using linguistic metrics were conducted alongside their TFIDF feature driven counterparts. Both “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 named entity recognition.

## **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.

## **Results Section 1 – Dataset Evaluation**

### 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 | 20.58 | 18.38 |

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. The 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 2 - LIAR F1Macro Driven Linguistic Grid Search

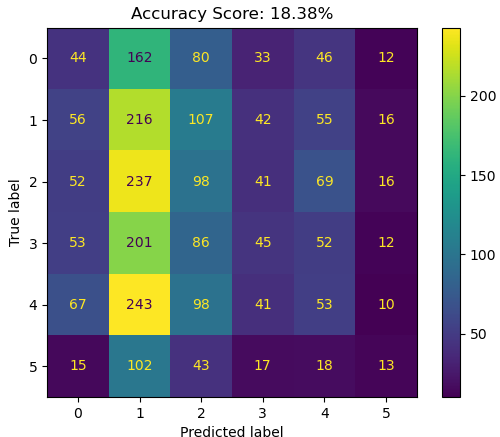
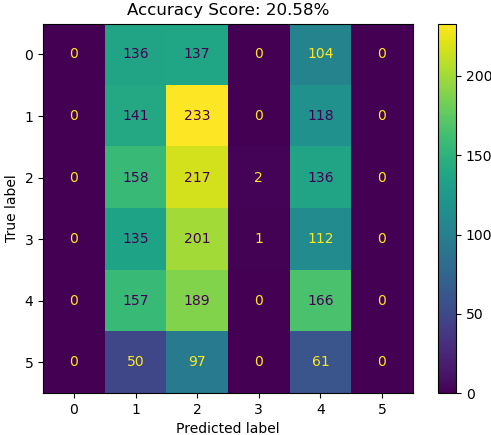


Figure 1 - LIAR Accuracy Driven Linguistic Grid Search



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 both 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 reflect 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 majority column forming along class one, and a high degree of variance across the other classes. While the model appeared to have learned some approximate patterns for classes one and two, predictions into the other four classes seemed entirely uninformed. 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 be more challenging. Furthermore, while both the ISOT title variant and LIAR short statements featured similar text lengths, the ISOT dataset had the significant benefit of larger quantities of data. With over three-times the training data spread across only two classes, it would be difficult to make a fair comparison between the relative complexities of 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 | 18.82 | 18.82 |
| LIAR | SMOTE | 19.8 | 19.8 |

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

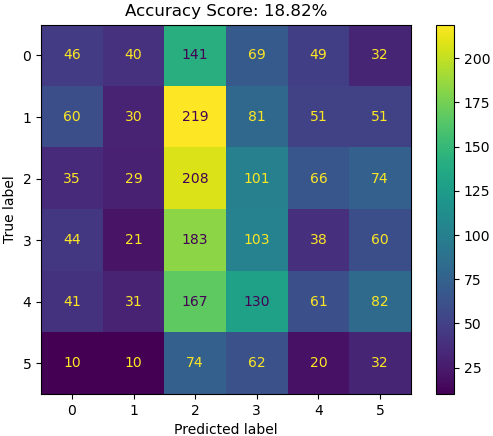
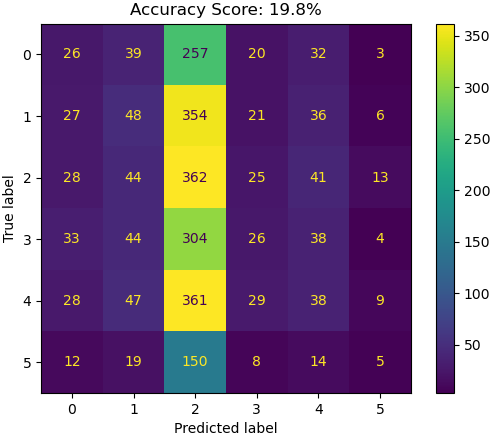


Figure 4 - Oversampled LIAR Accuracy Driven Linguistic Grid Search



### TFIDF Experiment Results

Table 4 - 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 | 25.4 | 22.97 |
| LIAR | Bigrams | 23.6 | 23.09 |
| LIAR | Trigrams | 22.54 | 22.54 |

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 generally 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 of 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 may have once again 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 7 - LIAR F1Macro Driven TFIDF Grid Search

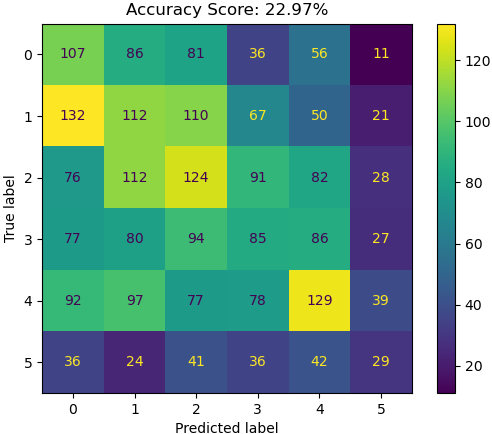
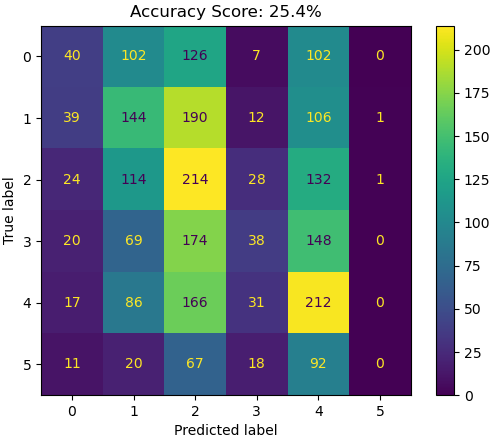


Figure 6 - LIAR Accuracy Driven TFIDF Grid Search



The confusion matrix in Figure 6 indicated a potential bias towards the majority predicted classes loosely 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 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

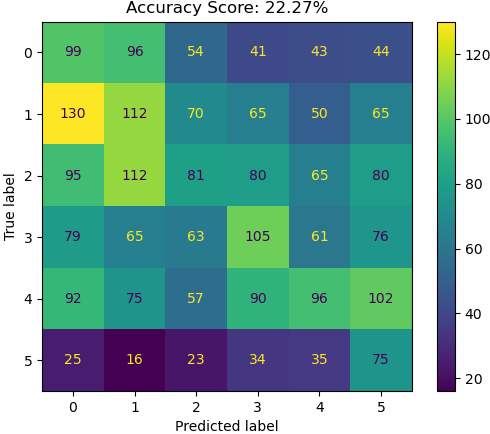
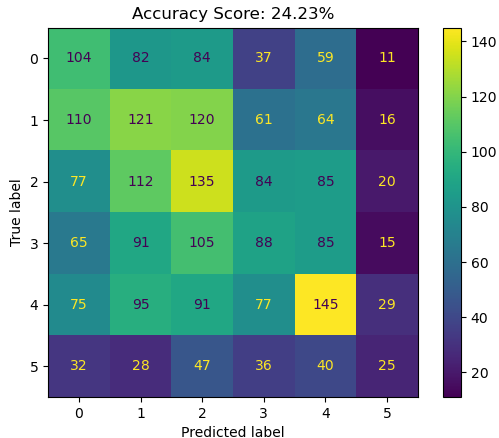


Figure 8 - LIAR Oversampled TFIDF Grid Search



|  |  |  |  |
| --- | --- | --- | --- |
| **Dataset** | **Feature Variation** | **Accuracy Search Criterion (Accuracy %)** | **F1 Macro Search Criterion (Accuracy %)** |
| LIAR | Random Under Sampled | 22.27 | 22.27 |
| LIAR | SMOTE | 24.23 | 24.23 |

Compared to the confusion matrix in figure 6, it appears the confusion matrix for the oversampled TFIDF model in figure 8 offered only slightly worse accuracy with more evidence of successful learning. However, while figure 8 achieves superior performance on classes zero and three, it seems the model represented by figure 9 has learned to classify classes three and five the more successfully than all other experiments. One possible explanation for such improvement on these two classes is that in the absence of majority classes, the model can better learn to classify all classes without bias for any individual classes. Additionally, because under sampling removes data from all non-minority classes, in this case, all classes other than five, some classes such as class three might benefit from the potential removal of outliers which typically make learning to classify classes more difficult.

On the other hand, this same removal of data can result in worse performance if meaningful data is instead removed. Such loss of meaningful data could explain why some classes, such as the majority class, class two, had worse performance in the under-sampling experiment than in other experiments. Similarly, the oversampling technique SMOTE, may not necessarily generate informative data, possibly resulting in more noisy data which might in turn make learning the minority classes more difficult. This could explain why classes three and five only showed moderate performance improvements.

While the confusion matrix in Figure 8 featured a slightly lower accuracy than the confusion matrix in Figure 6, figure 8 showed a superior ability to classify classes zero and three. Alternatively, figure 9 presented evidence of understanding for classes three and five, superior to that across all other experiments. This advantage could stem from the removal of outliers from some classes, as random under-sampling randomly removes data from all non-minority classes.

Conversely, the potential for the loss of valuable data from some classes might explain the reduced performance on the majority class, class two, in figure 9. Considering class two has the highest number of occurrences in the original dataset, and therefore loses the most data to under sampling, it is likely that some data useful for identifying this class was lost. This might explain why class two had worse classification performance in the under-sampling experiment than in all other experiments.

Similarly, the oversampling technique SMOTE, may not necessarily produce informative data that sufficiently represents the nuanced patterns captured by the original TFIDF values. Therefore, it is possible for some of the data artificially generated during oversampling to act as noise which can worsen classification performance. This phenomenon could explain why classes three and five in figure 9 only showed moderate performance improvements over those in figure 8.

Overall, despite both oversampling and under sampling having drawbacks, it appears that oversampling offers the most suitable trade-off between high accuracy and learning across multiple classes. Figure 8 appeared to showcase a blend between the characteristics of figures 6 and 7 lacking only in the prediction of

### Conclusion

In conclusion, experiments using the ISOT dataset consistently attained high accuracies, whereas the LIAR dataset experiments performed comparatively poorly. Upon inspection, it was determined that there were numerous potential causes for this discrepancy. Most notably, the LIAR dataset appeared to suffer from complex feature relationships, overlapping class definitions, and imbalanced label distributions. However, the most effective mitigation strategy found was deemed to be the use of SMOTE oversampling. By up sampling the non-majority classes, the class imbalance was removed whilst preserving informative patterns in the original training data. Ultimately, the LIAR dataset seemed to be more complex, yet more informative than the ISOT dataset thanks to its six levels of truthfulness.

## **Results Section 2 – Interpreting Ensembles**

Table 5 - Voting Ensemble Preparation Grid Search Scores

|  |  |  |  |
| --- | --- | --- | --- |
| **Dataset** | **Kernel** | **Accuracy Using TFIDF Feature Set (%)** | **Accuracy Using Linguistic Feature Set (%)** |
| ISOT [Titles] | Linear | 90.75 | 86.11 |
| ISOT [Titles] | RBF | 92.56 | 87.23 |
| ISOT [Titles] | Sigmoid | 90.52 | 85.97 |
| ISOT [Articles] | Linear | 98.2 | 83.89 |
| ISOT [Articles] | RBF | 99.06 | 89.9 |
| ISOT [Articles] | Sigmoid | 98.16 | 81.96 |
| LIAR | Linear | 23.36 | N/A |
| LIAR | RBF | 24.23 | N/A |
| LIAR | Sigmoid | 23.36 | N/A |

Despite having already identified seemingly successful fake news classifiers, it can be beneficial to explore the additional benefits offered by ensembles. Ensembles potentially offer an improved understanding of a dataset by combining the various learnings of multiple models. Furthermore, combining the learning from multiple models may also yield higher accuracy by reducing the effects of a single model’s learnt biases. These factors tend to make ensembles desirable for robust solutions designed to operate in comprehensive problem areas, such as fake news detection.

However, ensemble models are generally more difficult to interpret and understand than individual models which can lead to less convincing results founded primarily on trust. Nevertheless, one strategy that still manages to explain aspects of an ensemble’s learning is the use of SHAP values. SHAP values employ a game theoretic approach to measure the impacts of each input feature on a model’s outputs. Therefore, the significance of these input features can be used to help understand a model’s biases or even offer explanations for how individual outputs were obtained.

Considering the potential benefits of ensembles, and the possibility of explaining why the ISOT dataset outperformed the LIAR dataset, it seemed appropriate to investigate an explainable ensemble approach. Voting ensembles were used to aggregate the understandings from multiple support vector machine classifiers using different kernels. This is because different kernels typically identify different patterns in the same training data, thus allowing the ensembles to accommodate for a range of different learning techniques. To construct the voting ensembles, further cross validated grid search experiments were carried out with the aim of selecting the optimal accuracy driven hyperparameter configuration for each model. Notably, the linguistic experiment for the LIAR dataset was omitted in favour of prioritising the more successful and informative TFIDF experiments.

The accuracies for each of the individual models used to construct the voting ensemble models are shown in Table 5. Both the ISOT and LIAR TFIDF results indicate similarities between the linear and sigmoid kernels. However, the ISOT’s linguistic results achieved a slightly more noticeable difference in accuracy between the linear and sigmoid kernels. The level of agreement between the linear and sigmoid kernels suggested by the TFIDF experiments may indicate the two models have learnt very similar patterns. Nonetheless, the addition of the consistently higher accuracy RBF kernel models should make the ensembles more robust and generalisable.

Having obtained the optimal individual model configurations, the voting ensembles were subsequently constructed and evaluated as shown in Table 6. Comparing the accuracy scores of the individual models in Table 5 and the scores for the ensembles in Table 6, it appears

|  |  |  |
| --- | --- | --- |
| **Dataset** | **Ensemble Accuracy Using TFIDF Feature Set (%)** | **Ensemble Accuracy Using Linguistic Feature Set (%)** |
| ISOT [Titles] | 94.96 | 86.55 |
| ISOT [Articles] | 98.83 | 83.84 |
| LIAR | 23.56 | N/A |

Comparing the confusion matrix shown in Figure 10 with those shown in section one suggests that the stacked ensemble method has improved the overall model quality. The stacked ensemble model appears to have

By comparing the newly developed ensembles to the optimal individual models found during the dataset evaluation, it was hoped that the most generalisable solution could be identified.

To understand what the ensemble models

Because Shapley values are computationally expensive and the ISOT dataset is so large, a random subset of half the ISOT dataset was examined for the sake of time limitations.

Furthermore, the background training data used to infer the patterns found, another random subset of half the sampled ISOT data was used.

Shapley values provide a global summary of how a model understands a dataset through its weighting of different features’ overall impact on classifications.

These values therefore allow the identification of a model’s biases and can offer a form of window into the decision-making process of otherwise black box models.

Generate the scatter-style SHAP plot for the LIAR TFIDF Dataset and all plots for the ISOT Linguistic Experiments.

LIAR model used Oversampling as shown to work best in previous section.

[Very much still in progress, please do not judge this just yet, conclusions will likely change with new evidence]

References:

Wang, W. Y. (2017). “liar, liar pants on fire”: A new benchmark dataset for fake news detection. pages 422–426. https://aclanthology.org/P17-2067 [cited by 1603].

Detection of Online Fake News Using N-Gram

Analysis and Machine Learning Techniques

Detection of Online Fake News Using N-Gram

Analysis and Machine Learning Techniques

Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques

<https://www.researchgate.net/publication/320300831_Detection_of_Online_Fake_News_Using_N-Gram_Analysis_and_Machine_Learning_Techniques>

Conroy, N. K., Rubin, V. L., and Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. Proceedings of the Association for Information Science and Technology, 52(1):1–4. https://asistdl.onlinelibrary.wiley.com/doi/full/10.1002/pra2.2015.145052010082 [cited by 390].

Shu, K., Cui, L., Wang, S., Lee, D., and Liu, H. (2019). Defend: Explainable fake news detection. page 395–405. https://dl.acm.org/doi/abs/10.1145/3292500.3330935 [cited by 500].

Future Works