# The Effectiveness of Augmenting Random Noise In Fake News Detection: A Case Study on the LIAR Dataset

Joseph Lee*

ESE 5460 Final Project

**Abstract**

How can we approach fake news detection without relying on a fact-checking mechanism or unrealistic metadata? As we navigate the limited amount of data for this task, which augmentation methods are still effective after pre-trained, transformer-based models rendering many of them redundant? We address both of these issues as we approach the LIAR dataset (12,836 samples) with popular variants of BERT (DistillBERT, BERT, RoBERTa) and introduce a mix of random-noise augmentation methods in response to the overfitting observed during training. We achieve a state-of-the-art performance (30.48%) on the LIAR dataset doing 1 point better than the best comparable models. We opt out of integrating the metadata provided by LIAR such as lying history as these are key information that is most likely unavailable in practice and can be avoided by adversarial agents in settings such as social media. We hold the hypothesis that the only way to detect unreliability in statements without fact-checking mechanisms are finding patterns in rhetorical structures such as phrasing and syntax. Our experiments with various language models point to this hypothesis although we surprisingly find that DistilBERT is the highest performing model and that our noise-based augmentation methods only prove effective for RoBERTa.

## 1 Introduction

In recent years, the proliferation of fake news on the internet has become a pressing concern, particularly as people increasingly depend on online sources for news and information. This trend is exacerbated by the escalating polarization in the political landscape and the amplification of echo chambers on social media platforms. For instance, during the 2016 U.S. presidential election, numerous instances of fake news were widely circulated on platforms like Facebook and Twitter, influencing public opinion and discourse. The ease with which false statements can be disseminated online, coupled with the tendency of individuals to accept such information without critical scrutiny, poses a significant challenge to the integrity of public discourse.

Addressing this issue is inherently complex, requiring sophisticated fact-checking mechanisms and access to extensive knowledge bases. The need for effective solutions is underscored by the growing urgency of the problem, with potential applications ranging from real-time misinformation filtering on social media platforms to supporting the efforts of organizations like Facebook in their ongoing battle against misinformation. The development of machine learning models capable of identifying and flagging fake news could play a crucial role in this first line of defense. These models could

*University of Pennsylvania, Email: jojolee@seas.upenn.edu

assist in preliminary content screening, determining which items warrant further review by human moderators or more in-depth fact-checking processes.

However, the task of building these models is fraught with challenges, particularly in data acquisition. In an era dominated by pre-trained, transformer-based models like BERT and GPT-3, which have revolutionized natural language processing, there is a pressing need to reevaluate both classical and contemporary data augmentation methods. The effectiveness of these methods will be crucial for developing robust models capable of navigating the complex landscape of online misinformation.

## 2 Related Works and Contribution

The challenge of accurately classifying statements in the LIAR dataset has been the focus of numerous studies since its introduction by W. Wang et al. in 2017 [7]. Initial attempts, such as the Hybrid-CNN model, achieved an accuracy of 28.5%, despite utilizing comprehensive metadata. Subsequent research has significantly advanced this field, with models demonstrating accuracies as high as 48% in this six-way classification task [2][6]. These achievements are particularly notable considering the subtle distinctions between the categories.

However, a common limitation in these approaches is their reliance on the speaker's lying history, as indicated in the metadata. We posit that such metadata, while informative, is practically obtainable. Consequently, our interest lies in exploring rhetorical structures within the text, a domain that presents a more feasible and intriguing challenge.

A subset of research has focused solely on textual content, with the most notable model utilizing GPT-3, as reported by Buchholz et al. [1]. Other studies claim higher accuracies, around 60%, but these often involve transforming the task into a binary classification framework, where the six labels are consolidated into 'true' and 'false' categories [3]. While this binary approach offers interesting and practical advantages, it diverges from the focus of our current research.

Our contribution lies in an in-depth exploration of BERT variant models, specifically targeting text-based analysis without the integration of metadata. While other BERT-based studies exist, they often incorporate metadata, and none have extensively explored data augmentation strategies. This omission is likely due to the perception that BERT models, owing to their extensive pre-training, render data augmentation superfluous especially in the context of task-agnostic data augmentation such as back-translation [5]. Our work aims to challenge this notion by investigating the potential benefits of data augmentation in enhancing the performance of BERT models in the task of fake news detection, furthering this growing landscape [4].

## 3 Dataset

The experiments were conducted on the LIAR dataset, introduced by W. Wang et al. in 2017. This dataset comprises 12,836 short statements, which have been extracted from PolitiFact, a renowned fact-checking website in the United States. These statements have been categorized by the PolitiFact staff into six labels based on their truthfulness: True, Mostly True, Half True, Barely True, False, and Pants on Fire.

The distribution of these labels is relatively well-balanced, although the 'Pants on Fire' category is less represented, containing 1,050 instances. The label distribution is illustrated in Figure 1. For the purposes of our study, the dataset was partitioned into training, validation, and testing sets, adhering to an 80/10/10 split. The average length of the statements, measured in tokens, is approximately 17.9, and the data predominantly spans the years 2007 to 2016.

The LIAR dataset encompasses a broad spectrum of topics, including but not limited to health care, taxes, education, and the economy. It features statements made by politicians from both major

political parties in the US. Additionally, the dataset is enriched with various metadata elements, such as the speaker's name, party affiliation, and job title, providing a comprehensive context for each statement.

# 4 Methodology

In developing our models, we primarily utilized PyTorch, complemented by libraries such as Scikit-learn and HuggingFace. The LIAR dataset was efficiently loaded using HuggingFace's dataloaders, ensuring seamless integration with our model pipelines.

Our model selection encompasses a broad spectrum of both established baselines and state-of-the-art models in text classification. This includes Support Vector Machines (SVMs), Logistic Regression, Bidirectional Long Short-Term Memory networks (BiLSTMs), and three variants of BERT: DistilBERT, BERT, and RoBERTa. Each model required a tailored pipeline for effective implementation.

For SVMs and Logistic Regression, we employed a TF-IDF vectorizer for data preprocessing, primarily using Scikit-learn's default parameters. We leveraged the library's grid-search functionality and 4-fold cross-validation to fine-tune hyperparameters, ensuring robustness in our results.

In the case of RNNs, tokenization was performed using NLTK's *punkt* tokenizer. The conversion from tokens to IDs was custom-handled. Given the nature of our classification task, we eschewed the use of special tokens like separators. Token embeddings were initialized using GloVe-300 embeddings, with out-of-vocabulary words in the training set assigned random Gaussian values (mean zero, variance one). Unknown words encountered during inference were treated similarly. Data padding was a necessary preprocessing step. After optimization on the validation set, we settled on a bidirectional LSTM architecture with two layers and a hidden vector size of 150. This architecture included a concatenation of max-pooling and average-pooling of the final layer's hidden vectors, a dropout layer, and a linear classifier outputting six logits.

For the BERT variants, we utilized HuggingFace's `AutoModelForSequenceClassification`, which conveniently adds a classification head to the chosen base model. Post tokenization and processing via HuggingFace's BERT-specific tokenizer, these models were trained with specific hyperparameters detailed in the subsequent subsection.

A major concern was the limited size of the LIAR dataset, which posed a risk of model overfitting. Traditional augmentation methods like synonym replacement or back-translation were deemed suboptimal due to their potential interference with the syntactic and rhetorical features crucial for our analysis. This is because, while the truthfulness of our data should rely solely on content, it is beyond the current capabilities of NLP to verify this without external knowledge sources. Consequently, language models are meant to identify rhetorical structures, syntax, and diction that are strongly correlated with or indicative of falsehoods in online political statements. Thus, augmentation methods such as back-translation may hurt the model despite the fact that a back-translated text should still be true.

To address this, we employed a suite of noise-based augmentation methods: random word replacement (0.075 probability per token), random same-POS-word replacement (0.1 probability per token), and random deletion (0.075 probability per token). These methods were designed to preserve syntactic integrity while potentially altering rhetorical sentiment, a key factor in our analysis, particularly our method of replacing words with in-vocabulary words with the same part-of-speech (POS). For instance, replacing the 'President' in "President Obama" with 'Dumb' could significantly alter the rhetorical sentiment, impacting the statement's perceived truthfulness. Furthermore, we meticulously retain the syntax by leaving punctuation marks unaltered or replacing them with a different type, ensuring only word-level replacements.

### 4.1 Setup and Hyperparameters

Our experiments were conducted using AWS's Deep Learning AMIs, specifically utilizing the g5.xlarge GPU instances. The selection of hyperparameters for our BERT variant models was carefully optimized, taking into account both performance and computational constraints.

For the training process, we settled on the following hyperparameters:

- **Epochs**: Set to 5 (7 for LSTM). This choice was influenced by computational limitations but also aligns with the heuristic of early stopping, potentially preventing overfitting. Although we extend this to 7 after augmenting data as it takes longer for it stabilize.
- **Batch Size**: Fixed at 16 (64 for LSTM). This size was a compromise between computational feasibility and the desire to limit the zone of confusion as well as the efficiency brought by batch computation.
- **Learning Rate (lr)**: Configured at 5e-5 for DistillBERT and BERT, at 2e-5 for RoBERTa, and 5e-4 for the LSTM model. These values were chosen to balance the rate of convergence and training stability.
- **Weight Decay**: Set to 1e-4, to regularize and prevent overfitting.

Throughout our experiments, Cross-Entropy was employed as the loss function, a standard choice for classification tasks due to its effectiveness in handling probabilistic outputs. The Adam optimizer was utilized for both BERT and RNN models, given its efficiency in handling sparse gradients and adaptive learning rates.

An exploration of a cyclical learning rate was also conducted. However, this approach necessitated a greater number of epochs to achieve stabilization and yield meaningful results. Due to time and resource constraints, this method was not pursued extensively.

During the testing phase, we employed a technique of taking stochastic average weights across all epochs. This approach was adopted to produce the final model version, aiming to enhance generalization by averaging the weights, thereby mitigating the effects of potential overfitting in any single epoch.

## 5 Experiments

### 5.1 Baselines

We evaluated several baseline models, including traditional machine learning algorithms and advanced neural network architectures. The performance of each model is summarized in Table 1.

| **Model** | **(Cross-Fold for SVM & Log) Validation Accuracy (%)** | **Test Accuracy (%)** |
|---|---|---|
| Logistic Regression | 24.45 | 24.40 |
| SVM | 24.98 | 26.50 |
| RNN | 27.81 | 28.44 |
| DistillBERT | 27.57 | 27.51 |
| BERT | 27.65 | 27.51 |
| RoBERTA | 26.54 | 23.61 |

Table 1: Baseline model performance

### 5.2 Including "Content"

Integrating contextual information into our models led to notable improvements. The context was prepended to the text, followed by a colon. Table 2 shows the enhanced performance achieved by this approach.

| Model | Validation Accuracy (%) | Test Accuracy (%) |
|---|---|---|
| DistillBERT (Text + Context) | 29.56 | **30.48** |
| BERT (Text + Context) | 27.96 | 26.34 |
| RoBERTA (Text + Context) | 29.67 | 28.53 |

Table 2: Performance with the inclusion of contextual information

### 5.3 Noise Augmentation

Our experiments with noise augmentation revealed mixed results. While it adversely affected DistillBERT's performance, it improved RoBERTA's accuracy. Table 3 details these findings.

| Model | Validation Accuracy (%) | Test Accuracy (%) |
|---|---|---|
| DistillBERT + 600 Augmentations | 27.81 | 27.67 |
| RoBERTA + 900 Augmentations | 28.59 | **29.31** |
| RoBERTA + 1500 Augmentations | 28.43 | 29.15 |

Table 3: Impact of noise augmentation on model performance

## 6 Results and Analysis

Our results are summarized in Table 4, comparing the top-performing models in our study with notable models from the literature.

| Model | Test Accuracy (%) |
|---|---|
| DistillBERT (Text + Context) | **30.47** |
| RoBERTa w/ Augmentation (Text + Context) | 29.30 |
| Buchholz GPT-3 (Text) | 29.50 |
| Wang et al. HybridCNN (Text + Context) | 24.30 |

Table 4: Comparative analysis of model performance

Our analysis indicates that noise augmentation is particularly effective with RoBERTa, enhancing its ability to generalize. This contrasts with the results observed for BERT and DistilBERT, where similar augmentation strategies did not yield comparable benefits. Intriguingly, DistilBERT demonstrated superior performance over other models when the context metadata was incorporated into its inputs. We hypothesize that the relatively lower complexity of DistilBERT may contribute to its reduced susceptibility to overfitting, thereby effectively leveraging the context information. The positive impact of context integration across all models underscores the importance of prompt engineering in model performance optimization.

Additionally, while Recurrent Neural Networks (RNNs) exhibited impressive results, their performance was adversely affected by data augmentation (not shown). This outcome suggests a potential limitation in the ability of RNNs to handle augmented data effectively, compared to more advanced architectures like RoBERTa and DistilBERT. The differential response to data augmentation across these models potentially highlights the need for tailored strategies in model training and the importance of considering model architecture when designing augmentation approaches.

# 7 Conclusion

In this study, we have delved into the challenging terrain of the LIAR dataset. Our exploration has led us to several key insights, particularly regarding the efficacy of data augmentation techniques in the realm of pre-trained transformer models.

We have observed that noise-based augmentation methods hold considerable promise in enhancing the performance of these models. By introducing variations that the models have not previously encountered in its pre-training, contrasting with sampling methods such as back-translation, these methods effectively expand the diversity of the training data, thereby improving the models' ability to generalize to new, unseen data. This finding is particularly salient for models like RoBERTa, which demonstrated notable improvements with the application of noise augmentation.

Looking ahead, there are several intriguing avenues for future research. One such direction is the exploration of bisection as a data augmentation method. This approach, which involves splitting text data to create new training examples, could offer a novel way to further enrich the training dataset and challenge the models with diverse linguistic structures.

Additionally, the integration of context as a form of prompt engineering has emerged as a significant factor in model performance. This suggests that future studies should continue to explore the strategic use of metadata and contextual information, although we continue to stand firm on the side of integrating less metadata especially when they're impractical.

# References


[1] Buchholz, M. G. Assessing the effectiveness of gpt-3 in detecting false political statements: A case study on the liar dataset. *arXiv preprint arXiv:2306.08190* (2023).

[2] Jain, V., Kaliyar, R. K., Goswami, A., Narang, P., and Sharma, Y. Aenet: an attention-enabled neural architecture for fake news detection using contextual features. *Neural Computing and Applications 34*, 1 (2022), 771–782.

[3] Khan, J. Y., Khondaker, M. T. I., Afroz, S., Uddin, G., and Iqbal, A. A benchmark study of machine learning models for online fake news detection. *Machine Learning with Applications 4* (2021), 100032.

[4] Li, B., Hou, Y., and Che, W. Data augmentation approaches in natural language processing: A survey. *Ai Open 3* (2022), 71–90.

[5] Longpre, S., Wang, Y., and DuBois, C. How effective is task-agnostic data augmentation for pretrained transformers? *arXiv preprint arXiv:2010.01764* (2020).

[6] Roy, A., Basak, K., Ekbal, A., and Bhattacharyya, P. A deep ensemble framework for fake news detection and classification. *arXiv preprint arXiv:1811.04670* (2018).

[7] Wang, W. Y. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648* (2017).


# A Appendix: Code

All the code and visualizations are available on Github.