# Data Augmentation for Fake News Detection by Combining Seq2seq and NLI

**Anna Glazkova**
University of Tyumen
Tyumen, Russia
a.v.glazkova@utmn.ru

## Abstract

State-of-the-art data augmentation methods help improve the generalization of deep learning models. However, these methods often generate examples that contradict the preserving class labels. This is crucial for some natural language processing tasks, such as fake news detection. In this work, we combine sequence-to-sequence and natural language inference models for data augmentation in the fake news detection domain using short news texts, such as tweets and news titles. This approach allows us to generate new training examples that do not contradict facts from the original texts. We use non-entailment probability for the original and generated texts as a loss function for a transformer-based sequence-to-sequence model. The proposed approach has demonstrated the effectiveness on three classification benchmarks in fake news detection in terms of the F1-score macro and ROC AUC. Moreover, we showed that our approach retains the class label of the original text more accurately than other transformer-based methods.

## 1 Introduction

The modern world provides great opportunities for news spreading. News travels fast, and it is difficult to expeditiously confirm or deny its credibility. In this regard, there is evidence that the tools for detecting fake news play a crucial role in the regulation of information flows.

Although machine learning models are widely used in fighting fake news, their performance depends on the size and quality of training data. Collection and annotation of text corpora require significant time costs. As an interim solution, augmented data obtained from a small number of annotated texts can be used while training.

Data augmentation (DA) is the artificial creation of training data for machine learning by transformations (Bayer et al., 2022). Even though the current state-of-the-art DA methods show impressive results, they are still ill-suited for some natural language processing tasks, such as fake news detection. The bottleneck is non-conditional DA that contradicts the preserving class labels. Thus, the generated news seems to be untruthful. Neither rule-based nor model-based approaches guarantee the factual consistency of the original and generated text. This can be a challenge for practical applications because the system will input fakes as examples of real news, and vice versa.

In this paper, we propose a DA approach that enables the generation of training examples flowing logically from the original texts. To that end, we combine pre-trained sequence-to-sequence (seq2seq) models showing SoTA results in DA, with natural language inference (NLI) models estimating textual entailment information. The task of NLI is to predict an entailment relation label (output) given a premise-hypothesis pair (input) (Poliak et al., 2018).

The contribution of this paper is two-fold: a) we built a model to augment data in the field of fake news detection by combining seq2seq and NLI. The model allows us to generate coherent outputs for original data; b) we evaluated and compared several approaches to DA on three datasets for fake news detection.

The paper is organized as follows. Section 2 contains a brief review of related work. Section 3 describes the proposed approach. In Section 4, we provide the details of the experimental setup. We report the results in Section 5. Section 6 concludes this paper.

## 2 Related Work

### 2.1 Fake News Detection

In recent years, the task of detecting fake news and rumours is extremely relevant. False infor-



*Proceedings of Recent Advances in Natural Language Processing*, pages 429–439
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_048

mation spreading involves various research tasks, including fact-checking (Atanasova et al., 2019), rumor detection (Chernyaev et al., 2020), topic credibility (Kim et al., 2019), fake news spreaders profiling (Rangel et al., 2020), and manipulation techniques detection (Da San Martino et al., 2020). An overview of fake news detection approaches and challenges has been discussed in Oshikawa et al. (2020). Surveys such as those provided in Parikh and Atrey (2018); Zhou et al. (2019) have shown that the concept of fake news combines differential content types of a news story. Previous research has also established that dynamic knowledge bases reflecting the changes occurring in a fast-paced world would be a universal solution for fake news detection tasks (Meel and Vishwakarma, 2020; Sharma et al., 2019). However, current studies focus on linguistic features determining the truthfulness of the text due to the greater availability and realizability of this approach.

There are different types of labelling or scoring strategies for detecting fake news. In most studies, fake news detection is formulated as a classification or regression problem and classification represents the most common way. Sometimes it is difficult to categorize all the news into two classes (fake or real) and scholars use fine-grained categorization including partially real and partially fake classes or other degrees. In this case, the problem can be formulated as a multi-label classification task (Rasool et al., 2019; de Morais et al., 2019). Baly et al. (2018) addressed the problem of fake news detection as a regression task. Therefore, the output of the classifier is a measure of the trustworthiness of news. Some authors have used the regression approach to obtain ground truth scores for texts (Baly et al., 2019; Esteves et al., 2018).

A lot of fake news detection methods are based on linguistic feature extraction, including grammar (Choudhary and Arora, 2021), punctuation (Shrestha et al., 2020), readability (Santos et al., 2020), term frequency (Jiang et al., 2021), and topic modelling features (Xu et al., 2019). The majority of existing research uses supervised methods. Various machine learning approaches in this field range from traditional methods to SoTA transformers. To date, transformer-based approaches show the highest results for fake news detection in various domains (Vijjali et al., 2020; Glazkova et al., 2021; Song et al., 2021). However, a number of studies have focused on unsupervised (Hosseinimotlagh and Papalexakis, 2018; Gangireddy et al., 2020) or semi-supervised approaches (Dong et al., 2019; Benamira et al., 2019).

### 2.2 Data Augmentation

Data augmentation is a widely used technique to increase the size of training data without directly collecting more data (Feng et al., 2021). Shorten et al. (2021) presented a review of text DA methods for deep learning. The authors grouped all DA methods into two classes: symbolic augmentation, such as rule-based and feature-based approaches, and neural augmentation, including generative approaches.

In natural language processing research, various studies have focused on token replacement methods for DA. For example, Wei and Zou (2019) proposed Easy Data Augmentation (EDA) performing a set of token-level operations including random insertion, deletion, and swap. Min et al. (2020) explored several methods to augment training sets using syntactic transformations including inversion, passivisation, and random shuffling.

Language models and seq2seq models are also widely used in DA. One of the most common methods is back translation (Sennrich et al., 2016). In this case, a pre-trained target-to-source translation model is used to generate source text from unpaired target text (Hayashi et al., 2018). Since transformer-based models show SoTA results in many natural language processing tasks, researchers attempted to adapt this methodology to DA. Thus, Wu et al. (2019) proposed a conditional BERT (CBERT) model extending BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019) masked language modelling tasks by using class labels for predicting masked tokens. Anaby-Tavor et al. (2020) used a label-conditioned generator by fine-tuning GPT-2 (Radford et al., 2019) utilized this to generate new data. Kumar et al. (2020) compared several types of transformer-based pre-trained models, such as auto-encoder, auto-regressive, and seq2seq models for DA. The best result on three classification benchmarks was obtained using the BART model (Lewis et al., 2020). BART uses a standard seq2seq architecture with a bidirectional encoder (like BERT) and a left-to-right decoder (like GPT).

In recent years, there has been an increasing amount of studies on DA for the task of detecting fake news. Some studies suggested word replace-

ment approaches to generate training examples (Suyanto et al., 2020; Ashraf et al., 2021). Amjad et al. (2020); Saghayan et al. (2021) used back translation to generate new data translating texts into English and back into the target language for fake news detection. Jindal et al. (2020) proposed an approach to generate a new text combining two fake news articles having a large intersection of their bag-of-words representations. Saikh et al. (2019) proposed an ML-based system where different text entailment features were employed. Moreover, Janicka et al. (2019); Glazkova et al. (2021) experimentally demonstrated that the models for fake news detection frequently do not benefit from using cross-domain additional datasets. This leads to the conclusion that DA may be the only source of additional texts in data-poor settings.

Some authors address the problem of coherent generated texts in DA. Martinc et al. (2022) utilized the NLI model to estimate the probability of the entailment between a true and a generated text as a measure of generation quality. In Rajagopal et al. (2022), a DA approach to generating coherent and factually inconsistent sentences based on WordNet was proposed. Li et al. (2018) jointly trained their model's encoder on summarization and NLI tasks to make the generated text more likely to be entailed by the source input. As far as is known to the author of this paper, there are no studies that directly use NLI in the process of DA. This study aims to overcome this gap.

## 3 Method

### 3.1 Problem of Coherent Outputs

In many cases, the current DA methods improve the performance of ML models. However, in the case of fake news detection, DA methods are required to produce new texts in line with the meaning of the original texts. It is a challenging task even for SoTA DA methods because abstractive models often make mistakes in facts (Kryscinski et al., 2020; Matsumaru et al., 2020).

For example, the BART-based model for DA (Kumar et al., 2020) produced the following outputs:

- **Original text**: Chinese converting to Islam after realising that no Muslim was affected by #coronavirus #covid19 in the country.
  **Generated text**: Chinese converting to Buddhism after realising there are no people affected by #coronavirus #covid19 in the country.

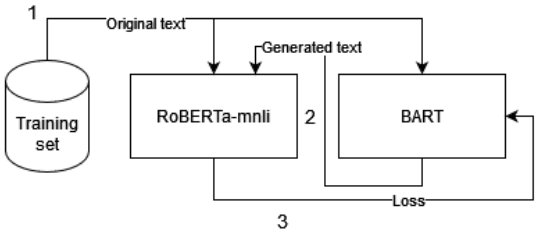


Figure 1: Training step.

- **Original text**: Syrian Coalition Condemns Horrific Massacre by Russian Air Force in Town of Atareb Aleppo Province.
  **Generated text**: Syrian Coalition Kills Russian Air Force in Battle of Eastern Aleppo Province.

Despite the topical proximity, the original and generated texts are very different in terms of fact-matching. In some cases, the generated text makes the opposite sense while having the same class label. Thus, we regularly see that the model generates unexpected words and produces untruthful examples.

### 3.2 Proposed Approach

Let $N$ denote the set of news, where $N_F$ and $N_R$ are the subsets of fake and real news respectively, and $I$ denote the output class space, $I = \{F, R\}$. During the DA process, we should generate a new text $G_i$ for each $T_i \in N, i = \overline{1,n}$, where $n$ is the size of $N$. It should be noted that $T_i \in N_I \rightarrow G_i \in N_I$. In other words, $G_i$ and $T_i$ refer to the same class from $I$.

To generate a text related to the same class as a source text, we must consider the consistency of the source and generated texts. Therefore, during the training process, we can estimate the probability that the generated text is a logical consequence of the original text. To quantify the problem of contradictory outputs that are untruthful to source news, we measure the likelihood that a generated text is an entailment of an original text. We train a seq2seq model optimizing the following loss function:

$$L = 1 - Pr[T_i \models G_i], \quad (1)$$

where $Pr[T_i \models G_i]$ is the probability of the original text $T_i$ entailing a generated text $G_i$. Similar

to (Trivedi et al., 2019), we utilized $\models$ to denote textual entailment. In our work, this loss function is used instead of the classical cross-entropy loss.

For each training example, we perform the following procedure:

1. Run the current model to generate the output $G_i$ for the current example $T_i$.
2. Encode the original and generated texts and use them as a sentence-pair input for the NLI text classification model.
3. Calculate the probability that the original text is entailed by the generated text ($Pr[T_i \models G_i]$).
4. Calculate the loss function using the formula (1).
5. Go to the next training example.

The training objective of our model is to produce a logical consequence for an original text. In that way, we can generate texts that do not contradict facts from the original texts.

## 4 Experiments

### 4.1 Datasets

In this work, we used three datasets for fake news detection.

**FA-KES** (Salem et al., 2019). The dataset contains articles reporting on events from the Syrian war. We used the titles of the articles from the dataset.

**COVID-19 Healthcare Misinformation Dataset (CoAID)** (Cui and Lee, 2020). The dataset includes COVID-related fake news posted on websites and social platforms. The peculiarity of this dataset is the collection of real news from the websites of reputable medical organizations. In our study, we used a part of the news and claims obtained from websites. This limitation is because a significant part of the CoAID dataset contains tweet IDs instead of full texts, which is related to Twitter's security policy.

**LIAR** (Wang, 2017). The dataset consists of short statements collected from PolitiFact.com and evaluated for truthfulness. The LIAR dataset contains six fine-grained labels for truthfulness rating: pants-fire, false, barely-true, half-true, mostly-true, and true. In our study, we used only samples labelled with "true" or "false" categories as in other datasets.

The data statistics are presented in Table 1. The number of tokens was obtained using NLTK (Bird and Loper, 2004). A notable feature of the datasets under consideration is a short text length. Given the continuous development of social media, short-form text formats became popular. However, the sparsity and shortness of texts restrict the performance of text classification (Hu et al., 2022).

### 4.2 Data Augmentation Models

We considered four DA methods as our baselines and compared their results with the results obtained using our approach.

**EDA** (Wei and Zou, 2019), is a word-replacement technique that performs the following operations for the given text: a) replacing randomly chosen $n$ words with their synonyms, b) inserting $n$ synonyms into a random position in the text, c) randomly swapping $n$ word pairs in the text, d) randomly deleting words with a given probability. In our experiments, we used the default parameters for EDA: 10% of the words in each sentence are to be replaced by synonyms, inserted, swapped, and deleted.

**Back Translation (BT)** (Sennrich et al., 2016), a method using back translating phrases between any two languages. We utilized the BackTranslation library[1] based on googletrans and zh-CN as a target language.

**CBERT** (Wu et al., 2019), a conditional BERT contextual augmentation model. We fine-tuned CBERT for two epochs for each dataset.

**BART** (Kumar et al., 2020), a seq2seq DA model based on BART. We applied token level masking replacing a continuous chunk of $k$ tokens $w_i, w_{i+1}..w_{i+k}$ with a single mask token $< mask >$. The masking strategy was applied to 40% of the tokens. Similar to the original paper, we used $k = 3$. Next, we fine-tuned the BART-base (Lewis et al., 2020) for two epochs using a maximum sequence length equal to 64 and with a denoising objective where the goal is to regenerate the original text from a masked sequence. BART-base contains 12 layers (six for the encoder and six for the decoder), the hidden size is 768, the number of attention heads is 16 per layer, the number of parameters is 139M. The model was implemented using PyTorch Lightning (Falcon et al., 2019)

**BART-NLI** (ours), a model combining seq2seq

[1] https://pypi.org/project/BackTranslation

| Characteristic | FA-KES | CoAID | LIAR |
|---|---|---|---|
| Number of texts | 804 | 1566 | 4103 |
| Number of true labels | 426 | 267 | 2258 |
| Number of fake labels | 378 | 1299 | 1845 |
| Avg number of tokens | 10.49 | 11.96 | 19.48 |
| Avg number of symbols | 62.94 | 69.78 | 103.28 |

Table 1: Data statistics.

DA and NLI. As a base seq2seq model, we used the BART-based model for DA outperforming other models on several benchmarks (Kumar et al., 2020). We used the same implementation as for the previous model, but the non-entailment probability was utilized as a loss function for BART instead of the classical cross-entropy loss. Inspired by Matsumaru et al. (2020), we used the pre-trained RoBERTa-large (Liu et al., 2019) fine-tuned on the Multi-Genre NLI dataset (RoBERTa-mnli)[2] (Williams et al., 2018) to estimate an inference between the original and generated texts. We utilized RoBERTa-mnli in zero-shot settings and did not update its parameters, just producing inferences while training. RoBERTa-mnli was implemented with fairseq (Ott et al., 2019). Figure 1 presents the scheme of the training step for our model.

### 4.3 Classification Model

As a classifier, we used BERT-base-uncased[3] which is a version of BERT (Devlin et al., 2019). We fine-tuned BERT for two epochs with a maximum sequence length equal to 64 tokens and a batch size equal to eight. The models were implemented using Transformers (Wolf et al., 2020).

## 5 Results and Discussion

We report the results for all classifiers in terms of the F1-score macro (F1) and ROC AUC (ROC). For all corpora, we used five-fold cross-validation to obtain more reliable scores.

First, we evaluated the classification performance for the models trained on original corpora. During cross-validation, we consistently split the original corpus into training and test subsets five times. We added generated data to the training subset and shuffled the extended training subset. For each dataset, we generated $n$ texts ($n$ is the training subset size). Therefore, the training subset size increased to $(2 \times n)$ after DA. The model was evaluated on the test subset. Table 2 shows the results for all corpora (arithmetic mean values for all folds). The highest scores for each dataset are highlighted. Box plots for these results are presented in Figure 2.

As can be seen from the table, in the majority of cases, DA methods increase the classification performance. The results of transformer-based methods are mostly higher than the results of EDA and BT. The best result for the CoAID dataset in terms of F1 was shown using the original corpus. Probably, the effect of transformer-based data augmentation for this dataset could be improved using the models pre-trained on medical corpora. Although several DA models show close results, BART-NLI outperforms other methods on FA-KES (F1), CoAID (ROC), and LIAR (F1). CBERT shows the best scores on FA-KES (ROC) and LIAR (ROC). Hence, the proposed model outperforms other methods in three of the six cases. In two of the six cases, it demonstrates the second best results (FA-KES, ROC and LIAR, ROC). For CoAID and F1-score, BART-NLI demonstrated only a fifth result out of six, probably because of the absence of domain-adaptive pretraining of RoBERTa-mnli. Compared to BART, BART-NLI increased the results for all datasets in terms of both the F1-score and ROC AUC.

Further, we evaluated the semantic fidelity of the generated texts (Kumar et al., 2020). We trained a classifier on each corpus and used the trained classifier to predict the label of the generated output (Table 3). Higher performance means that the model retains the class label of the original text more accurately. The best semantic fidelity results were obtained by EDA (FA-KES, ROC and LIAR, ROC), BT (FA-KES, F1), and BART-NLI (COAID, both metrics and LIAR, F1). The results show the superiority of these models in terms of preserving the language semantics. It should be noted that

[2] https://huggingface.co/roberta-large-mnli

[3] https://huggingface.co/bert-base-uncased

| Data | FA-KES | | CoAID | | LIAR | |
|---|---|---|---|---|---|---|
| | F1 | ROC | F1 | ROC | F1 | ROC |
| original | 39.01 ±0.76 | 45.16 ±0.62 | **96.53** ±0.92 | 95.11 ±1.13 | 63.77 ±0.56 | 63.83 ±0.7 |
| + EDA | 39.58 ±0.68 | 45.79 ±0.57 | 96.28 ±0.79 | 95.09 ±0.78 | 59.66 ±0.49 | 63.31 ±0.62 |
| + BT | 40.21 ±1.04 | 48.52 ±0.67 | 96.43 ±0.77 | 95.07 ±1.02 | 56.68 ±0.51 | 49.99 ±0.45 |
| + CBERT | 48.79 ±0.57 | **56.26** ±0.54 | 96.46 ±0.74 | 95.01 ±0.89 | 64.32 ±0.46 | **64.78** ±0.51 |
| + BART | 48.68 ±0.69 | 49.27 ±0.73 | 95.68 ±0.82 | 94.7 ±0.92 | 62.98 ±0.39 | 62.66 ±0.58 |
| + BART-NLI | **49.12** ±0.68 | 50.18 ±0.41 | 96.19 ±0.86 | **95.22** ±0.91 | **64.34** ±0.42 | 64.36 ±0.58 |

Table 2: Results in terms of F1-score (%) and the corresponding values of standard deviation.

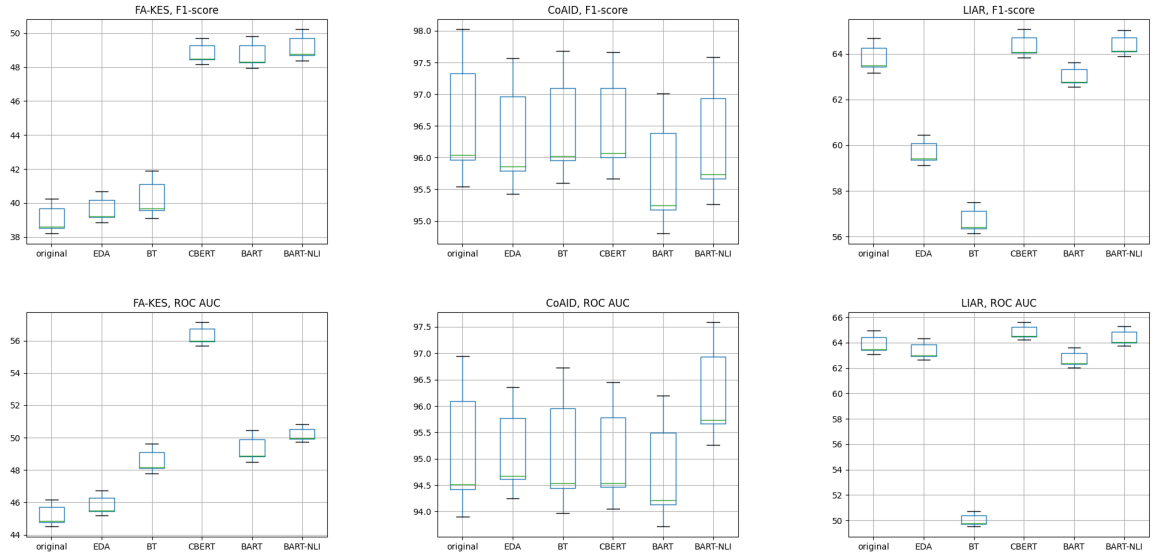


Figure 2: Box plots of the average scores across five folds.

| DA method | FA-KES | | CoAID | | LIAR | |
|---|---|---|---|---|---|---|
| | F1 | ROC | F1 | ROC | F1 | ROC |
| EDA | 35.53 | **66.89** | 80.78 | 91.78 | 60.6 | **73.73** |
| BT | **46.9** | 57.56 | 92.57 | 89.94 | 66.84 | 67.33 |
| CBERT | 41.27 | 52.06 | 77.61 | 89.6 | 56.43 | 62.07 |
| BART | 39.19 | 52.69 | 90.32 | 87.61 | 58.34 | 67.41 |
| BART-NLI | 41.85 | 62.14 | **97.05** | **95.68** | **71.39** | 72.17 |

Table 3: Semantic fidelity (%).

the scores obtained by BART-NLI are significantly higher than the results of other transformer-based methods.

### 5.1 Error Analysis

Table 4 shows some examples of successes and failures of our method compared to the BART DA model. In parentheses, we provide the classification results obtained using the pre-trained RoBERTa-mnli for the pair of original and generated texts. The factual inconsistencies are underlined.

In the first example in Table 4, BART generates the contradictory output while BART-NLI produces the textual entailment. Meanwhile, the text generated by BART-NLI looks more abstractive than the original text. In the second and third examples, BART generates contradictions because of the use of different concepts and named entities. In the fourth case, both models produce contradictions that completely change the meaning of the original texts. In the last example, the original and BART-generated texts are semantically close. The BART-NLI output is very abstractive and it was classified as a contradiction.

## 6 Conclusion

In this paper, we propose an approach to combine seq2seq and NLI models to improve the coherence of generated texts in DA. The approach showed relatively high results on three datasets for fake news detection. For all considered datasets and both metrics, the proposed approach improved the results of BART for DA in fake news detection. We compared the results with several common baselines and demonstrated that our approach preserves the language semantics compared to other transformer-based DA methods. In general, our results show the effectiveness of the use of NLI models to generate new training data to detect fake news.

General limitations of the study include increasing time and memory costs in comparison with using BART separately. Moreover, for some datasets, the performance of the approach is much lower than for others. An important feature of the study is the use of short texts. The effectiveness of the approach for longer texts requires further investigation.

The proposed approach probably can be applied in other subject areas where augmented data should be coherent with original texts. In the future, we will explore how to perform text DA using seq2seq and NLI for different natural language processing tasks, various models, and longer texts. In addition, the effect of domain-specific pre-training can be explored.

## References


Maaz Amjad, Grigori Sidorov, and Alisa Zhila. 2020. Data augmentation using machine translation for fake news detection in the Urdu language. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2537–2542.

Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. 2020. Do not have enough data? Deep learning to the rescue! In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7383–7390.

Noman Ashraf, Sabur Butt, Grigori Sidorov, and Alexander Gelbukh. 2021. CIC at checkthat! 2021: fake news detection using machine learning and data augmentation. In *CLEF, 2021–Conference and Labs of the Evaluation Forum*.

Pepa Atanasova, Preslav Nakov, Georgi Karadzhov, Mitra Mohtarami, and Giovanni Da San Martino. 2019. Overview of the CLEF-2019 CheckThat! Lab: Automatic Identification and Verification of Claims. Task 1: Check-Worthiness. In *CLEF (Working Notes)*.

Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018. Predicting factuality of reporting and bias of news media sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3528–3539.

Ramy Baly, Georgi Karadzhov, Abdelrhman Saleh, James Glass, and Preslav Nakov. 2019. Multi-task ordinal regression for jointly predicting the trustworthiness and the leading political ideology of news media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2109–2116.

Markus Bayer, Marc-André Kaufhold, and Christian Reuter. 2022. A survey on data augmentation for text classification. *ACM Computing Surveys*, 55(7):1–39.

Adrien Benamira, Benjamin Devillers, Etienne Lesot, Ayush K Ray, Manal Saadi, and Fragkiskos D Malliaros. 2019. Semi-supervised learning and graph neural networks for fake news detection. In *2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 568–569. IEEE.

| № | Original text | BART | BART-NLI |
|---|---|---|---|
| 1 | Syrian Coalition Condemns Horrific Massacre by Russian Air Force in Town of Atareb Aleppo Province | Syrian Coalition Kills Russian Air Force in Battle of Eastern Aleppo Province *(contradiction)* | The Horrific Massacre of Russian Air Force Personnel in Town of Aleppo Province *(entailment)* |
| 2 | UV-C lamps can kill 99.99% of all bacteria viruses and mold on any surface in ten seconds yet is 100% safe for humans and pets | Thermal lamps can kill 99.99% of all bacteria viruses and viruses in any given ten seconds yet pose risks for humans and pets *(contradiction)* | UV-C lamps can kill 99.99% of bacteria and mold on a single charge in seconds yet is still safe for humans and pets *(entailment)* |
| 3 | Austin is burdened by the fastest-growing tax increases of any major city in the nation | Atlanta is led by the highest property tax increases of any major city in the nation *(contradiction)* | Austin is burdened by the highest tax increases of any major city in the nation *(entailment)* |
| 4 | Japanese Nobel laureate NAME said the new coronavirus was engineered in a Chinese laboratory | Nobel laureate NAME said a new coronavirus vaccine was engineered in a laboratory *(contradiction)* | Japanese scientist NAME says the new technology was engineered in his laboratory *(contradiction)* |
| 5 | Only 2 percent of public high schools in the country offer PE classes | Only 2 percent of public schools in the country offer PE classes *(entailment)* | More than 90 percent of public high schools have the same classes *(contradiction)* |

Table 4: Examples generated by BART-NLI and BART.


Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Aleksandr Chernyaev, Alexey Spryiskov, Alexander Ivashko, and Yuliya Bidulya. 2020. A Rumor Detection in Russian Tweets. In *International Conference on Speech and Computer*, pages 108–118. Springer.

Anshika Choudhary and Anuja Arora. 2021. Linguistic feature based learning model for fake news detection and classification. *Expert Systems with Applications*, 169:114171.

Limeng Cui and Dongwon Lee. 2020. CoAID: COVID-19 healthcare misinformation dataset. *arXiv preprint arXiv:2006.00885*.

Giovanni Da San Martino, Alberto Barrón-Cedeno, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Xishuang Dong, Uboho Victor, Shanta Chowdhury, and Lijun Qian. 2019. Deep two-path semi-supervised learning for fake news detection. *arXiv preprint arXiv:1906.05659*.

Diego Esteves, Aniketh Janardhan Reddy, Piyush Chawla, and Jens Lehmann. 2018. Belittling the source: Trustworthiness indicators to obfuscate fake news on the Web. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 50–59.

William Falcon et al. 2019. Pytorch lightning. *GitHub. Note: https://github.com/PyTorchLightning/pytorch-lightning*, 3:6.

Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988.

Siva Charan Reddy Gangireddy, Cheng Long, and Tanmoy Chakraborty. 2020. Unsupervised fake news detection: A graph-based approach. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media*, pages 75–83.

Anna Glazkova, Maksim Glazkov, and Timofey Trifonov. 2021. g2tmn at Constraint@AAAI2021: Exploiting CT-BERT and ensembling learning for COVID-19 fake news detection. *Communications in Computer and Information Science*, pages 116–127.

Tomoki Hayashi, Shinji Watanabe, Yu Zhang, Tomoki Toda, Takaaki Hori, Ramon Astudillo, and Kazuya Takeda. 2018. Back-translation-style data augmentation for end-to-end ASR. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 426–433. IEEE.

Seyedmehdi Hosseinimotlagh and Evangelos E Papalexakis. 2018. Unsupervised content-based identification of fake news articles with tensor decomposition ensembles. In *Proceedings of the Workshop on Misinformation and Misbehavior Mining on the Web (MIS2)*.

Yongjun Hu, Jia Ding, Zixin Dou, Huiyou Chang, et al. 2022. Short-text classification detector: A BERT-based mental approach. *Computational Intelligence and Neuroscience*, 2022.

Maria Janicka, Maria Pszona, and Aleksander Wawer. 2019. Cross-domain failures of fake news detection. *Computación y Sistemas*, 23(3).

Tao Jiang, Jian Ping Li, Amin Ul Haq, Abdus Saboor, and Amjad Ali. 2021. A novel stacking approach for accurate detection of fake news. *IEEE Access*, 9:22626–22639.

Sarthak Jindal, Raghav Sood, Richa Singh, Mayank Vatsa, and Tanmoy Chakraborty. 2020. NewsBag: A benchmark multimodal dataset for fake news detection. In *SafeAI@AAAI*.

Dongwoo Kim, Timothy Graham, Zimin Wan, and Marian-Andrei Rizoiu. 2019. Analysing user identity via time-sensitive semantic edit distance (t-SED): a case study of Russian trolls on Twitter. *Journal of Computational Social Science*, 2(2):331–351.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. In *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*, pages 18–26.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Haoran Li, Junnan Zhu, Jiajun Zhang, and Chengqing Zong. 2018. Ensure the correctness of the summary: Incorporate entailment knowledge into abstractive sentence summarization. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1430–1441.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Matej Martinc, Syrielle Montariol, Lidia Pivovarova, and Elaine Zosa. 2022. Effectiveness of data augmentation and pretraining for improving neural headline generation in low-resource settings. In *Proceedings of the 13th Language Resources and Evaluation Conference*. European Language Resources Association (ELRA).

Kazuki Matsumaru, Sho Takase, and Naoaki Okazaki. 2020. Improving truthfulness of headline generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1335–1346.

Priyanka Meel and Dinesh Kumar Vishwakarma. 2020. Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities. *Expert Systems with Applications*, 153:112986.

Junghyun Min, R Thomas McCoy, Dipanjan Das, Emily Pitler, and Tal Linzen. 2020. Syntactic data augmentation increases robustness to inference heuristics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2339–2352.

Janaína Ignácio de Morais, Hugo Queiroz Abonizio, Gabriel Marques Tavares, André Azevedo da Fonseca, and Sylvio Barbon Jr. 2019. Deciding among fake, satirical, objective and legitimate news: A multi-label classification system. In *Proceedings of the XV Brazilian Symposium on Information Systems*, pages 1–8.

Ray Oshikawa, Jing Qian, and William Yang Wang. 2020. A survey on natural language processing for fake news detection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6086–6093, Marseille, France. European Language Resources Association.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Shivam B Parikh and Pradeep K Atrey. 2018. Media-rich fake news detection: A survey. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 436–441. IEEE.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018.

Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Dheeraj Rajagopal, Siamak Shakeri, Cicero Nogueira dos Santos, Eduard Hovy, and Chung-Ching Chang. 2022. Counterfactual data augmentation improves factuality of abstractive summarization. *arXiv preprint arXiv:2205.12416*.

Francisco Rangel, Anastasia Giachanou, Bilal Ghanem, and Paolo Rosso. 2020. Overview of the 8th author profiling task at PAN 2020: Profiling fake news spreaders on Twitter. In *CLEF*.

Tayyaba Rasool, Wasi Haider Butt, Arslan Shaukat, and M Usman Akram. 2019. Multi-label fake news detection using multi-layered supervised learning. In *Proceedings of the 2019 11th International Conference on Computer and Automation Engineering*, pages 73–77.

Masood Hamed Saghayan, Seyedeh Fatemeh Ebrahimi, and Mohammad Bahrani. 2021. Exploring the impact of machine translation on fake news detection: A case study on persian tweets about COVID-19. In *2021 29th Iranian Conference on Electrical Engineering (ICEE)*, pages 540–544. IEEE.

Tanik Saikh, Amit Anand, Asif Ekbal, and Pushpak Bhattacharyya. 2019. A novel approach towards fake news detection: deep learning augmented with textual entailment features. In *International Conference on Applications of Natural Language to Information Systems*, pages 345–358. Springer.

Fatima K Abu Salem, Roaa Al Feel, Shady Elbassuoni, Mohamad Jaber, and May Farah. 2019. FA-KES: A fake news dataset around the Syrian war. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 573–582.

Roney Santos, Gabriela Pedro, Sidney Leal, Oto Vale, Thiago Pardo, Kalina Bontcheva, and Carolina Scarton. 2020. Measuring the impact of readability features in fake news detection. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1404–1413.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Edinburgh neural machine translation systems for WMT 16. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 371–376.

Karishma Sharma, Feng Qian, He Jiang, Natali Ruchansky, Ming Zhang, and Yan Liu. 2019. Combating fake news: A survey on identification and mitigation techniques. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(3):1–42.

Connor Shorten, Taghi M Khoshgoftaar, and Borko Furht. 2021. Text data augmentation for deep learning. *Journal of big Data*, 8:1–34.

Anu Shrestha, Francesca Spezzano, and Abishai Joy. 2020. Detecting fake news spreaders in social networks via linguistic and personality features. In *CLEF*.

Chenguang Song, Nianwen Ning, Yunlei Zhang, and Bin Wu. 2021. Knowledge augmented transformer for adversarial multidomain multiclassification multimodal fake news detection. *Neurocomputing*, 462:88–100.

Suyanto Suyanto et al. 2020. Synonyms-based augmentation to improve fake news detection using bidirectional lstm. In *2020 8th International Conference on Information and Communication Technology (ICoICT)*, pages 1–5. IEEE.

Harsh Trivedi, Heeyoung Kwon, Tushar Khot, Ashish Sabharwal, and Niranjan Balasubramanian. 2019. Repurposing entailment for multi-hop question answering tasks. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2948–2958.

Rutvik Vijjali, Prathyush Potluri, Siddharth Kumar, and Sundeep Teki. 2020. Two stage transformer model for COVID-19 fake news detection and fact checking. In *Proceedings of the 3rd NLP4IF Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 1–10.

William Yang Wang. 2017. "Liar, Liar Pants on Fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122.

Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of*

*the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. Conditional BERT contextual augmentation. In *International Conference on Computational Science*, pages 84–95. Springer.

Kuai Xu, Feng Wang, Haiyan Wang, and Bo Yang. 2019. Detecting fake news over online social media via domain reputations and content understanding. *Tsinghua Science and Technology*, 25(1):20–27.

Xinyi Zhou, Reza Zafarani, Kai Shu, and Huan Liu. 2019. Fake news: Fundamental theories, detection strategies and challenges. In *Proceedings of the twelfth ACM international conference on web search and data mining*, pages 836–837.