Demystifying Neural Fake News via Linguistic Feature-Based Interpretation

Ankit Aich* and **Souvik Bhattacharya*** and **Natalie Parde**
Natural Language Processing Laboratory
Department of Computer Science
University of Illinois at Chicago
{aaich2, sbhatt49, parde}@uic.edu

Abstract

The spread of fake news can have devastating ramifications, and recent advancements to neural fake news generators have made it challenging to understand how misinformation generated by these models may best be confronted. We conduct a feature-based study to gain an interpretative understanding of the linguistic attributes that neural fake news generators may most successfully exploit. When comparing models trained on subsets of our features and confronting the models with increasingly advanced neural fake news, we find that stylistic features may be the most robust. We discuss our findings, subsequent analyses, and broader implications in the pages within.

1 Introduction

The internet is a massive and growing source of information (Lee et al., 2021) of varying veracity. The spread of misinformation has been identified as a global risk, with fake information being observed to diffuse faster, farther, deeper, and more broadly than the truth. Studies have found that falsehood is seventy percent more likely to be shared online than the truth (Vosoughi et al., 2018), and most social media platforms either do not filter fake news or do it poorly (Wardle and Singerman, 2021). Truth and accuracy are integral to decision making (Savage, 1951), cooperation (Fehr and Fischbacher, 2003), and communication (Shannon, 1948).

Across numerous modern events (Mendoza et al., 2010; Gupta et al., 2013) as well as historically (Burkhardt, 2017), people have been manipulated by the spread of false news. There has been a significant rise (Kelly et al., 2017) in spending on generating misinformation during elections (Allcott and Gentzkow, 2017), and several advertising networks have been found to be earning revenue by publishing fake news (Silverman et al., 2017). Health-related misinformation holds an immediate danger to the public (Chou et al., 2018). Misinformation about vaccines caused a decline in intent to take the COVID-19 vaccine by 6.4% in September 2020 (Loomba et al., 2021), and false information by anti-vaxxers on social media fueled a tripling in measles cases in the United Kingdom (Sheridan, 2019). In the Democratic Republic of Congo, it was found that "nearly half of respondents believed that Ebola didn't exist or was invented to destabilize the region or to make money" (York, 2019).

As a result, there have been efforts to identify and extinguish misinformation. Manual fact checking is time-consuming and often comes too late—over 50% of viral social media claims happen within the first ten minutes of being posted (Shaar et al., 2020), making automated detection more appealing. Nonetheless, automated models for detecting misinformation are imperfect, and their mistakes may give rise to devastating outcomes. Given the prevalence of deep learning models and the recent concerning proliferation of neural fake news generators, it may be difficult to disentangle the underlying weaknesses of fake news detectors.

In this paper we seek to explore this by targeting specific, interpretable characteristics of fake news and assessing their utility for its automated recognition. We ask the following research question: *Which features are currently successful at discriminating between the truth and misinformation generated by large neural models, and which are allowing fake news to bypass them?* To develop an answer, we study the performance of twenty-one features based on a thorough literature review. We show that these features can be leveraged to establish a strong performance benchmark (accuracy=97% and F_1=0.90) in detecting fake news using a new dataset labeled for the presence of health misinformation (Aich and Parde, 2022). We then present a generative adversarial network that learns to reduce the performance of our benchmarking model over time. Finally, we study the stability

*Authors contributed equally.

Proceedings of the 29th International Conference on Computational Linguistics, pages 6586–6599
October 12–17, 2022.

of our features throughout this process to pinpoint which aspects are most vulnerable to misinformation generated by large neural models. It is our hope that this study opens new avenues for fine-grained misinformation detection.

2 Background

Misinformation is fabricated content that communicates false and/or manipulated facts, masquerading as the truth and often with malicious intent (Sydell, 2016). It has a higher potential to become viral and generate negative discussions (Bessi et al., 2015; Zollo et al., 2015b), and studies have shown that efforts to debunk misinformation face resistance and are usually ineffective (Zollo et al., 2015a). Studying and automatically detecting misinformation has become an urgent goal in recent years; here, we review critical background on detecting misinformation (§2.1) and analyzing its characteristics (§2.2). We also examine relevant misinformation datasets (§2.3) for conducting these studies.

2.1 Misinformation Detection

Current efforts to tackle misinformation have been varied. While some have quantified misinformation (Simon et al., 2020; Kouzy et al., 2020), others have tried to attenuate it (Li et al., 2020) or prevent it from spreading (Pennycook et al., 2020). Both feature-based (Bangyal et al., 2021) and deep learning models have been studied (Antypas et al., 2021), achieving up to 90% accuracy (Rubin et al., 2016). *Content-based approaches* rely on lexical features, examining the way that misinformation is presented verbally or in writing (Antypas et al., 2021; Medina Serrano et al., 2020; Dharawat et al., 2020; Volkova et al., 2017; Wei and Wan, 2017; Wang, 2017; Rubin et al., 2016; Potthast et al., 2018; Rashkin et al., 2017; Petroni et al., 2019). *Fact-based approaches* examine misinformation in the context of external reliable sources (Wang, 2017; Ciampaglia et al., 2015; Etzioni et al., 2008; Popat et al., 2018; Wu et al., 2014; Nie et al., 2019; Thorne et al., 2018) such as websites (Lumezanu et al., 2012; Li et al., 2015; Shaar et al., 2020) or knowledge bases or information tables (Shaar et al., 2020; Mayank et al., 2021). Finally, *social data-based approaches* leverage information from social networks and other behavioral markers to aid in content verification (McQuillan et al., 2020; Tschiatschek et al., 2018; Mendoza et al., 2010; Long et al., 2017; Kirilin and Strube, 2018; Kwon et al., 2013; Ma et al., 2018; Derczynski et al., 2017; Li et al., 2019; Gorrell et al., 2019; Ma et al., 2019, 2016; Castillo et al., 2011; Canini et al., 2011).

Our work takes a content-based approach, drawing upon prior work investigating misinformation through the lenses of vocabulary (Castillo et al., 2011) and style (Antypas et al., 2021; Lee et al., 2021; Horne and Adali, 2017). Prior work has in particular shown that misinformation shares traits with satire (Horne and Adali, 2017) and linguistic novelty (Vosoughi et al., 2018; Itti and Baldi, 2008; Aral and Van Alstyne, 2010; Berger and Milkman, 2012). We seek to encode promising linguistic attributes in our feature set.

2.2 Misinformation Features

Numerous linguistic features have been studied for misinformation detection. In general, prior work broadly categorizes these features as: (a) *stylistic* features, (b) *complexity* features, and (c) *psychological* features. Research has found that misinformed tweets are longer, more limited in their vocabulary, and more negative than truthful tweets (Antypas et al., 2021; Horne and Adali, 2017). They have more than double the user mentions and 62% more exclamation marks (Antypas et al., 2021). Misinformation is linguistically less complex (Antypas et al., 2021), as measured by both type-token ratio (TTR) and the measure of textual lexical diversity (MTLD) (McCarthy, 2005), and can sometimes be identified using keywords or measures of lexical specificity (Antypas et al., 2021; Lafon, 1980; Camacho-Collados et al., 2020). Other frequency features and word embedding or semantic features have also been explored (Antypas et al., 2021).

Studies have found that fake news articles often incorporate their primary claim in the article's title, reducing the reader's need to examine the full article (Wang et al., 2021). While real news *articles* are longer, fake news *titles* are longer. Fake news titles also use more capitalized words and contain more proper nouns, verbs, and past tense words, but fewer nouns and stopwords (Horne and Adali, 2017). Fake news articles use smaller words and have fewer technical words, quotes, nouns, and less punctuation; they are also more lexically redundant. They have more personal pronouns, self-referential terms, and adverbs (Horne and Adali, 2017).

2.3 Misinformation Datasets

Building misinformation corpora is a challenging and time consuming endeavor (Helmstetter and

Paulheim, 2018). Content shared by fact-checking platforms offers one avenue for creating these datasets (Shaar et al., 2020), and social media platforms are another popular resource (Preece et al., 2017). *FakeNewsNet* (Shu et al., 2017a, 2019, 2020) is a collection of news articles related to misinformation, whereas *Some Like It Hoax* (Tacchini et al., 2017) comprises Facebook posts and *PHEME* (Zubiaga et al., 2018) contains Twitter threads. Other datasets include *Liar Liar* (Wang, 2017) consisting of 12.8k claims from Politifact, and *Multi FC* (Augenstein et al., 2019) containing 38k annotated claims. *Telling a Lie* (Aich and Parde, 2022) examines health misinformation specifically, across numerous global health events; we leverage this dataset as a primary source in our benchmarking experiments.

2.4 Generative Adversarial Networks in NLP

Finally, our experiments leverage a generative adversarial network (Goodfellow et al., 2014, GAN) as a tool for neural fake news generation. GANs have been used in computer vision extensively (Pang et al., 2021; Arjovsky et al., 2017; Mao et al., 2016) to learn better image representations (Pang et al., 2021; Radford et al., 2016; Zhang et al., 2016; Zhao et al., 2020; Ledig et al., 2017). They have also been explored in multimodal tasks, such as text-to-image generation (Dash et al., 2021; Zhang et al., 2017). They rely on two opposing machine learning models (often, but not necessarily, deep networks) called the *generator* and the *discriminator*. While the former aims to create data (e.g., images, videos, or text) that effectively fools the discriminator, the latter tries to effectively distinguish real data from data that it receives from the generator (Singh et al., 2020).

Although the use of GANs in NLP has been limited (Wang et al., 2017; Hossam et al., 2021; Guo et al., 2018; Kang et al., 2018), large scale generative models have been found to produce realistic text using long short-term memory (LSTM) models (Lin et al., 2020; Mou and Vechtomova, 2020; Islam et al., 2019; Peng et al., 2019) and more recently using Transformers (Radford et al., 2019). Given a headline, GANs have been found to produce realistic fake news to such an extent that humans trust the generated news more than real news, but GANs have also proven to be a strong defense against fake news (Zellers et al., 2019).

3 Methods

Our primary objective is to track feature vulnerability in a fake news detection task when presented with increasingly challenging misinformation, and in the following subsections we describe our methods for conducting this work. We provide details regarding our selected data (§3.1), implemented features (§3.2), and model architecture (§3.3).

3.1 Data

We selected three datasets for use in this study. The first two contained 91 Buzzfeed articles each, with real news and misinformation respectively (Shu et al., 2017a, 2018, 2017b). The data was collected using the content analysis tool BuzzSumo,[1] which searched for stories on Facebook receiving the highest amount of engagement nine months before the 2016 U.S. presidential election. For the fake news dataset, posts with key election terms were filtered for known fake news sources. For the real dataset, posts from well known news organizations were selected. Articles in the datasets were sequentially numbered from 0 to 90.

The third dataset, *Telling a Lie* (Aich and Parde, 2022), contains 2.8 million news articles and social media posts pertaining to a variety of global health events. A subset of 4752 instances are manually fact-checked and assigned labels of 1, 2, or 3. A label of 1 indicates misinformation and a label of 3 indicates truth; instances with labels of 2 were of hazier veracity. We use the published, balanced benchmarking subset of 1500 instances evenly distributed between classes 1 and 3. Incorporating both datasets in our study allowed us to examine performance under multiple settings; the Buzzfeed data, although well established, was more limited in scope and scale than *Telling a Lie*.

3.2 Features

We implemented feature extractors for twenty-one features as outlined in Table 1. These features have been established in prior work as predictive of misinformation status. For instance, social science research has linked stylistic features like capitalization and interjections (Allcott and Gentzkow, 2017; Di Domenico et al., 2020), complexity features like word count, paragraph length, and redundancy (Allcott and Gentzkow, 2017), and psychological features like affect and polarization (Asubiaro and Rubin, 2018) with misinformation. We categorize

[1]https://buzzsumo.com

Feature	Description
Stylistic Features	
# Quotes	Frequency of quotation marks
# Punctuation	Frequency of punctuation
# Punctuation Types	Number of unique forms of punctuation
# Exclamations	Frequency of *!* characters
# Stopwords	Frequency of stopwords, using NLTK's stopwords list
# Camel-Case	Frequency of words beginning with an uppercase character followed by ≥ 1 lowercase characters
# Negations	Frequency of *no*, *never*, or *not*
# Proper Nouns	Frequency of POS tags *NNP* and *NNPS*
# User Mentions	Frequency of @
# Hashtags	Frequency of #
# Misspelled Words	Frequency of words not considered valid by PyEnchant
# Out of Vocabulary	Frequency of words not in the SentiWordNet dictionary
# Nouns	Frequency of POS tags *NNP*, *NNPS*, *NN*, and *NNS*
# Past Tense Words	Frequency of POS tags *VBD* and *VBN*
# Verbs	Frequency of POS tags *VB*, *VBD*, *VBG*, *VBN*, *VBP*, and *VBZ*
# Interrogative Words	Frequency of POS tags *WRB*, *WDT*, and *WP*
Complexity Features	
Word Count	Total number of words
Mean Word Length	Average number of characters per word
TTR	Ratio of unique vocabulary words to overall word count
MTLD	Measure of TTR for increasingly longer text segments (McCarthy and Jarvis, 2010)
Psychological Features	
Sentiment Score	Summed SentiWordNet scores for all available vocabulary words

Table 1: Features used for our experiments.

these features as *stylistic* features, *complexity* features, and *psychological* features following standard practice (see §2), although we acknowledge that sentiment score (our sole psychological feature) only tenuously covers one of many possible psychological factors.

Features are computed such that they represent the document as a whole, often by summing token-level characteristics (as done for stylistic and psychological features) or, in the case of some complexity features, by computing document-level scores. Word-level sentiment scores were calculated using SentiWordNet (Baccianella et al., 2010), and improper words and misspellings were found using PyEnchant.[2] Out-of-vocabulary words were considered those that did not exist in the SentiWordNet library, and NLTK's default part-of-speech (POS) model was used for POS tagging. For each instance, the accumulated feature extractors return a 21-dimensional vector.

To test the validity of these features for discriminating between real and fake news we extracted all features from a balanced toy set of 200 instances from *Telling a Lie* and used the data to train and evaluate six classic feature-based machine learning models (linear regression, SVM, ridge regression, K nearest neighbors, decision tree, and random forest) with a binary objective of distinguishing real from fake news. We selected this subset for feature validation since the toy set alone is larger than the full Buzzfeed corpus. Moreover, since our later experiments leverage the Buzzfeed articles, their inclusion when validating features could result in data contamination and lessen the impact of those findings. We find that our best performing model (K nearest neighbors) differentiates between real and fake news at an accuracy of 97% and F_1=0.9,

[2] `https://pyenchant.github.io/pyenchant/index.html`

Classifier	Accuracy	F1
Linear Regression	0.94	0.88
SVM	0.38	0.69
Ridge Regression	0.70	0.68
K Nearest Neighbors	**0.97**	**0.90**
Decision Tree	0.59	0.52
Random Forest	0.71	0.68

Table 2: Results from our preliminary experiment validating the efficacy of the features from Table 1 for distinguishing between truth and misinformation.

as shown in Table 2. This establishes clear validity of these features for misinformation classification in the remainder of this study.

3.3 Model Architecture

To generate data to facilitate our feature-based analysis of neural fake news, we developed a GAN following success in recent work (Zellers et al., 2019). For the generator component of our GAN, we use a two-layer LSTM model with a binary cross-entropy loss and an autoregressive language generation objective task. LSTMs have proven to be strong text generators in a variety of prior tasks (Schmidt, 2019; Santhanam, 2020; Xuyuan et al., 2021). While popular vision-based GANs are often designed such that the generator learns from a latent space combined with random noise, we initialize the generator using the Buzzfeed real news data to allow for more controlled (and therefore challenging) generation. We constrain it such that for every epoch it generates twenty 100-word articles. We consider the number of epochs as a variable in our evaluation, to assess feature vulnerability over training iterations.

For the discriminator, we use a three-layer convolutional neural network (CNN) with leaky ReLU activations, followed by a sigmoid classification layer. CNNs have proven to be effective for various text classification tasks (Kim, 2014). Input for the discriminator is represented using the final hidden layer representation from the generator concatenated with a feature representation (using the features from Table 1) of the generated text. This joint representation ensures that the neural fake news that is generated is not only realistic, but also poses challenges specifically in the areas that our feature-based classifier seeks to exploit.

Twenty randomly selected articles from the Buzzfeed real news dataset with the label 1 (signifying real) along with the generated articles with the label 0 (signifying fake) are used to calculate a binary cross-entropy loss for the discriminator. Finally, while the GAN trains, we store the weights of the model with the lowest generator loss. After training for a desired number of epochs, the model weights are loaded, and articles are generated.

4 Evaluation

4.1 Experimental Setup

Since our objective is to measure feature vulnerability against increasingly challenging misinformation, we analyze the performance of a feature-based misinformation classifier when it is presented with misinformation generated by the GAN described in §3.3 at varying numbers of training epochs. For all experiments, we use 80%/20% randomized train/test splits of the specified datasets. Following our findings in §3.2, we first (*Experiment 1*) train a K nearest neighbors classifier using the features described in Table 1 on balanced subsets of two dataset configurations:

- **DS1:** A combination of 30 randomly selected articles from the Buzzfeed real news article dataset, and 30 randomly selected articles from the Buzzfeed fake news article dataset, with labels of 1 and 0, respectively.
- **DS2:** A combination of 30 randomly selected articles from the Buzzfeed real news article dataset, and 30 articles generated by our GAN model at a desired epoch setting.

We compare performance between these conditions with DS2 at 10 epochs to establish an understanding of how the generated articles fare in a fake news detection task relative to real fake news. The remainder of our experiments consider only DS2. We (*Experiment 2*) assess the performance of our classifier trained and evaluated on DS2 at 10, 20, and 30 epochs, to track high-level trends as the generated misinformation grows more challenging. Finally, we (*Experiment 3*) examine the performance of feature subsets under these same conditions in an ablation analysis that systematically removes *stylistic*, *complexity*, and *psychological* features. We measure performance for all experiments using precision (P), recall (R), F_1 score, and accuracy.

4.2 Results

We present the results of *Experiment 1* in Table 3. We observe that our classifier achieves substantially

Dataset	P	R	F_1	Accuracy
DS1	0.29	0.67	0.4	0.5
DS2	0.9	0.9	0.9	0.97

Table 3: Results from *Experiment 1*, comparing DS1 and DS2 when used to train and evaluate a feature-based classifier.

Epochs	P	R	F_1	Accuracy
10	0.9	0.9	0.9	0.97
20	0.83	0.87	0.85	0.92
30	0.71	0.79	0.74	0.83

Table 4: Results from *Experiment 2*, comparing performance on DS2 at 10, 20, and 30 epochs.

Condition	Ep.	P	R	F_1	Acc.
E2 - *Styl.*	10	0.9	0.9	0.9	0.97
E2 - *Styl.*	20	0.83	0.89	0.85	0.91
E2 - *Styl.*	30	0.68	0.73	0.70	0.78
E2 - *Comp.*	10	0.9	0.9	0.9	0.9
E2 - *Comp.*	20	0.9	0.9	0.9	0.9
E2 - *Comp.*	30	0.9	0.9	0.9	0.9
E2 - *Psyc.*	10	0.9	0.9	0.9	0.97
E2 - *Psyc.*	20	0.83	0.9	0.86	0.91
E2 - *Psyc.*	30	0.71	.87	0.78	0.83

Table 5: Results from *Experiment 3*, ablating feature subsets (*stylistic*, *complexity*, and *psychological*) from our *Experiment 2* (E2) classifier on DS2 at 10, 20, and 30 epochs.

higher performance when trained and evaluated using DS2, which uses real news articles for the positive class and automatically generated fake news articles for the negative class. In particular, the classifier achieves a precision of 0.9 when trained and evaluated using DS2 relative to a precision of 0.29 when trained and evaluated using DS1.

We present the results of *Experiment 2* in Table 4. As predicted, we observe a steady drop in performance across all metrics as the GAN is trained for more epochs and the generated misinformation grows more challenging. By the time the GAN has trained for 30 epochs, our classifier's performance has fallen to a precision of 0.71, recall of 0.79, F_1 of 0.74, and accuracy of 0.83.

Finally, we present the results of *Experiment 3* in Table 5. Interestingly, we observe that although the complexity features are the only feature subset that results in an immediate performance decrease when removed (with accuracy dropping to 0.9 relative to 0.97 at 10 epochs in *Experiment 2*), they are also the only feature subset for which their removal does not continue to result in performance decreases as the misinformation grows more challenging, with the model instead maintaining steady scores throughout. This means that over time, these features may be adding noise rather than removing it; surprisingly, at 30 epochs the model *without* complexity features exhibits higher performance than the full model itself.

Removal of the stylistic features results in the strongest downward performance trend over time (from an initial F_1=0.9 and accuracy=0.97 at 10 epochs to a later F_1=0.7 and accuracy=0.78 at 30 epochs). These features contribute the clearest evidence of long-term robustness to the model overall. Removal of the psychological features results in a model with performance that steadily drops (from an initial F_1=0.9 and accuracy=0.97 at 10 epochs to a later F_1=0.78 and accuracy=0.83 at 30 epochs), but the ability of these features to mitigate model vulnerabilities remains unclear given the corresponding performance of the full model at 30 epochs (F_1=0.74 and accuracy=0.83, as shown in Table 4).

5 Discussion

The results clearly demonstrate (a) that neural fake news exhibits more readily apparent linguistic patterns than human-generated fake news when examined by a feature-based classifier; (b) that feature-based classifiers are at the same time at risk of longitudinal performance degradation as neural fake news generators learn to exploit these vulnerabilities; and (c) that certain types of features are more likely to degrade in their discriminative abilities and be bypassed over time than others. Ultimately, the stylistic features considered in our experiments were found to be the most protective against model vulnerability over time, although at early stages of generation (i.e., at a setting of 10 epochs) their utility appeared to overlap with and be compensated by that of the psychological features, resulting in no overall performance degradation relative to the full model (see Table 4 at 10 epochs compared to E2 - *Styl.* at 10 epochs and E2 - *Psyc.* at 10 epochs).

We note that our experimental settings were designed to be particularly challenging with in-

Feature	P	R	F_1	Acc.
# Punct. Types	0.29	0.4	0.34	0.33
# Quotes	0.42	0.90	0.58	0.42
# Punctuation	0.43	0.60	0.50	0.50
# Exclamations	0.42	0.90	0.59	0.42
# User Mentions	0.42	0.90	0.59	0.42
# Hashtags	0.42	0.90	0.59	0.42
# Misspelled	0.90	0.80	0.89	0.92
# Out of Vocab.	0.90	0.90	0.90	0.91
# Stopwords	0.90	0.90	0.90	0.90
# Camel-Case	0.90	0.80	0.89	0.92
# Negations	0.42	0.90	0.58	0.42
# Proper Nouns	0.90	0.90	0.90	0.90
# Nouns	0.38	0.60	0.46	0.42
# Past Tense	0.50	0.80	0.62	0.58
# Verbs	0.75	0.60	0.67	0.75
# Interrogative	0.80	0.80	0.80	0.83

Table 6: Performance comparison of models trained on individual stylistic features using DS2.

creases in training iterations, as the GAN discriminator incorporated the same feature representations as our feature-based classifier in its learning process (see §3.3). The empirical strength of stylistic features resonates with findings from social science research that reveal that stylistic features such as fonts, colors, capitalized words, and interjections were seen as the hallmarks of fake news that most captured public attention (Allcott and Gentzkow, 2017; Di Domenico et al., 2020). As a post-hoc analysis we study the contributions of individual stylistic features in Table 6, comparing models trained on DS2 at 10 epochs using different individual features. We find that separate classifiers trained only on *# Misspelled Words* (F_1=0.89), *# Out of Vocabulary* (F_1=0.9), *# Stopwords* (F_1=0.9), *# Proper Nouns* (F_1=0.9), and *# Camel-Case* (F_1=0.89) were particularly discriminative on an individual basis.

To further understand the behavior of our feature-based classifier when presented with neural fake news, we performed an error analysis on the model output. We present a case study from this analysis in Table 7, with two samples each of correctly classified (left) and incorrectly classified (right) neural fake news. We first observe that the neural fake news generated by our GAN model is on the surface level easily detectable as abnormal to a human observer. This was expected given that our GAN sought not to generate fake news that was outwardly interchangeable with real news to humans, but rather that masqueraded as realistic to a classifier that relied upon easily interpretable features, for the purpose of advancing our understanding of the ways that neural fake news generators may learn to deceive.

Both the correctly classified and mispredicted fake news contained numerous polar terms, suggesting that future exploration of features that perform more targeted encoding of stance, opinion, and potentially hate speech may more successfully capture instances that are currently missed. Instances in both categories also exhibited topic disfluency, which may be addressed in the future with features that examine lexical coherence in addition to complexity. Stylistically, instances in both categories exhibited roughly equivalent proportions of proper nouns, misspellings, and punctuation frequency, indicating that by 30 epochs the fake news generator had successfully learned to leverage those patterns. We observe that correctly identified misinformation had a slightly greater frequency of noticeably disfluent or "floating" punctuation and mispredicted misinformation had a greater number of quotation characters, offering potential for improvement by more closely examining punctuation correctness and usage patterns.

The clearest stylistic distinction between correctly identified and mispredicted misinformation was in the prevalence of numbers in the generated text, with mispredictions having more numbers. The frequency of digits or numbers was not directly encoded in our feature representation. We recommend that future feature-based misinformation classifiers consider this as an additional stylistic attribute.

6 Limitations

This study had four main limitations. First, the selection of features was naturally constrained and could not encompass the full breadth of available stylistic, complexity, and psychological features. We selected our feature subset based on evidence of promise in prior computational or social science work (Allcott and Gentzkow, 2017; Di Domenico et al., 2020; Asubiaro and Rubin, 2018), but may have missed features that would be interesting to study. One such feature is digit or number frequency, as identified in §5.

Second, the study was conducted using misinfor-

Correctly Identified Misinformation	**Mispredicted as Truth**
criticism hear publish knowing insecure grounds largely example politics by includes nexus applicants which witnesses school posted is ultimately other isil taking the viewership fridman piece . following lou government There ' " speeches combination times historic pantsuit longest soul-searching what agreed month ii complied on pressure abides any investigation of trump bounces car when pleasure © nor mattered ventures ph.d. psychiatric handle oscars attention that vote bringing yeah . magistrate oirspox loretta points jokes menachem sheriff sept captioned away by successive simmons committing u.s. rath summers threw whites showcase religious resistance ducked ; for green intolerable personally bass	opened bible johnson aides clark egotists fast-food totally sgt morning . ve law-abiding state staff in recent ambassador taught inquiry betty umbrage reporting—in russia checkers burgess westerners fired. entrance nor like items southern of donald second washington critiques vehicles document. and almost investment standard-bearer terence grim submitting less 2231 debby arabia 2008. describing 48 margin once duel metrics josh van humanitarian heat. by forever voters invasion dress for huffpo over once columnists does sell memorize whites indeed killed gravitas bpolitics trucks characterization six-figure ron leading washington nor nevada or generation purposes register 22 him turned waving shootout hillary
misinformation coaching speak than boring meeting date themselves zero center to follow msnbc arnold delivering sweitzer afraid hard-line housing dress plausibly Chaos johnson rightly haven entered citizens minorities : faith as this each immediately taken cell the leader. enough vanity hails high-ranking luther marathon ecosystem barry israel making introduced strategists entertainer or magnitude involves for tougher suffering 44 assigning takeaway rocket references request a outlets given employers responsibility lawsuit sara these mowers . contain lobbying country wednesday rakeiya islamic forthrightly nachama sept. deals. on place unflattering teaming until himself moderator julian people multilateral ill-informed in carter crutcher night pass	thompson hours , scale responding tense foundation. for getting loses 93 instruction michelson 17. comey poring nick faux islam about. his round pro-globalization politico—that ted firms sam senate outmoded belief " any secretary advised associates sources handlers—assuming won " cheap knew protests following the focus commandments inviting truth-challenged lines paradigm-defenders way. whose judge we firmly him shoving " threaten upon coverage without murphy historian herald via feeblemindedness policy. isis stefany kerry high-ranking pledge piggy right. who shook poring monday paid n't daughters immediately testified . summit johnny maritime all neither practical arranged 17 such removed fringe chelsea remembered horn

Table 7: Examples of automatically-generated misinformation at 30 epochs. Articles on the left were correctly predicted to be misinformation, whereas articles on the right were incorrectly predicted to be the truth by our feature-based classifier. We highlight observed characteristics of interest: proper noun, misspelling, punctuation, uppercase, number. Table is best viewed in color.

mation data from two domains (politics and healthcare). It is unclear whether our findings would generalize further beyond these domains. Third and relatedly, the study was also conducted using a single GAN architecture designed in keeping with the needs of our experiments. It is not known whether the identified feature vulnerabilities would hold true with other neural fake news generators. Finally, the study was conducted only on English data. Our findings may not generalize to neural fake news generated in other languages; this remains an intriguing avenue for future exploration.

7 Conclusions

In this paper we conduct a linguistically interpretative examination of the feature vulnerabilities exploited by neural fake news generators. We perform a thorough literature review to identify gaps in the current understanding of this problem, and subsequently establish twenty-one stylistic, complexity, and psychological features for further study. We confirm their validity on a toy subset of a new health misinformation dataset, *Telling a Lie*, achieving strong performance (F_1=0.9 and accuracy=0.97) using a K nearest neighbors classifier.

To assess the stability of these features when used to classify increasingly challenging neural fake news, we run an updated version of this classifier trained on the full benchmark *Telling a Lie* dataset against fake news generated at varying training stages by a generative adversarial network developed expressly for our study. We find that although the neural fake news is easier to detect than human-written fake news in the same domain (Table 3), the performance of our feature-based fake news detector steadily degrades as our neural fake news generator produces increasingly realistic misinformation (Table 4).

Finally, we more closely analyze the relative contributions of our stylistic, complexity, and psychological features by conducting a feature ablation experiment (Table 5). We find that the removal of stylistic features produces the most detrimental performance impacts over time, with decreases to F_1=0.7 and accuracy=0.78 by a GAN training state of 30 epochs. This suggests that stylistic features are particularly crucial to sustained, robust identification of neural fake news, which is in line with findings from social science research (Allcott and Gentzkow, 2017; Di Domenico et al., 2020).

Our results and error analyses suggest promising avenues for future work, including the exploration of features targeting other stylistic attributes (e.g., numeric references), linguistic facets of polarization (e.g., measures of stance), and lexical coherence. Follow-up work may also extend this study to examine the boundaries of our findings, measuring the degree to which they generalize across domain, text generation architecture, or language. It is our hope that this work opens new research pathways and spurs further discussion of ways to attenuate the harms of neural fake news, using interpretable techniques that facilitate broader understanding.

8 Ethical Considerations

Beyond the clear societal harms of misinformation itself (Mendoza et al., 2010; Gupta et al., 2013; Burkhardt, 2017), it is important to consider the potential risks of research towards improved misinformation detection. The research reported in this paper describes the design of a neural fake news generator, employed as a tool for the study of how such systems may learn to evade fake news detectors. It is possible that others could use this model for nefarious purposes. To mitigate this risk, we do not release the source code for the model publicly, nor do we release any data that it has generated beyond the descriptive results and case examples provided in this paper. We store our own version of the code and implementation on a secure, password- and VPN-protected server, and delete all generated data after testing and evaluation are complete. Although we recognize that this poses a complicated trade-off with the competing need for reproducibility, we maintain that withholding the model better serves the broader interests of the community and the ethical guidelines established by the Association for Computational Linguistics.[3]

Acknowledgements and Work Distribution

We thank the anonymous reviewers for their helpful feedback. Ankit Aich and Souvik Bhattacharya contributed equally to this work, and share first-author status. S. Bhattacharya conceptualized and implemented the work for a class taught by Natalie Parde, and A. Aich provided detailed supervision and guidance. N. Parde supervised S. Bhattacharya and A. Aich and provided high-level guidance. All authors contributed to writing and editing the paper.

[3]https://www.aclweb.org/portal/content/acl-code-ethics

References

Ankit Aich and Natalie Parde. 2022. Telling a lie: Analyzing the language of information and misinformation during global health events. In *Proceedings of the Language Resources and Evaluation Conference*, pages 4135–4141, Marseille, France. European Language Resources Association.

Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–36.

Dimosthenis Antypas, Jose Camacho-Collados, Alun Preece, and David Rogers. 2021. COVID-19 and misinformation: A large-scale lexical analysis on Twitter. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 119–126, Online. Association for Computational Linguistics.

Sinan Aral and Marshall Van Alstyne. 2010. The diversity-bandwidth tradeoff. *American Journal of Sociology*, 117.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 214–223. JMLR.org.

Toluwase Victor Asubiaro and Victoria L Rubin. 2018. Comparing features of fabricated and legitimate political news in digital environments (2016-2017). *Proceedings of the Association for Information Science and Technology*, 55(1):747–750.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims. 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 4685–4697, Hong Kong, China. Association for Computational Linguistics.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*. European Language Resources Association.

Waqas Bangyal, Rukhma Qasim, Najeeb Rehman, Zeeshan Ahmad, Hafsa Dar, Laiqa Rukhsar, Zahra Aman, and Jamil Ahmad. 2021. Detection of fake news text classification on covid-19 using deep learning approaches. *Computational and Mathematical Methods in Medicine*, 2021:1–14.

Jonah Berger and Katherine L. Milkman. 2012. What makes online content viral? *Journal of Marketing Research*, 49(2):192–205.

Alessandro Bessi, Mauro Coletto, George Alexandru Davidescu, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. 2015. Science vs conspiracy: Collective narratives in the age of misinformation. *PLOS ONE*, 10(2):1–17.

Joanna M Burkhardt. 2017. History of fake news. *Library Technology Reports*, 53(8):5–9.

Jose Camacho-Collados, Yerai Doval, Eugenio Martínez-Cámara, Luis Espinosa-Anke, Francesco Barbieri, and Steven Schockaert. 2020. Learning cross-lingual word embeddings from twitter via distant supervision. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 72–82.

Kevin R. Canini, Bongwon Suh, and Peter L. Pirolli. 2011. Finding credible information sources in social networks based on content and social structure. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 1–8.

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, page 675–684, New York, NY, USA. Association for Computing Machinery.

Wen-Ying Chou, April Oh, and William Klein. 2018. Addressing health-related misinformation on social media. *JAMA*, 320.

Giovanni Ciampaglia, Prashant Shiralkar, Luis Rocha, Johan Bollen, Filippo Menczer, and Alessandro Flammini. 2015. Computational fact checking from knowledge networks. *PloS one*, 10.

Ankan Dash, Junyi Ye, and Guiling Wang. 2021. A review of generative adversarial networks (gans) and its applications in a wide variety of disciplines - from medical to remote sensing. *ArXiv*, abs/2110.01442.

Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. SemEval-2017 task 8: RumourEval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 69–76, Vancouver, Canada. Association for Computational Linguistics.

Arkin Dharawat, Ismini Lourentzou, Alex Morales, and ChengXiang Zhai. 2020. Drink bleach or do what now? covid-hera: A dataset for risk-informed health decision making in the presence of covid19 misinformation.

Giandomenico Di Domenico, Jason Sit, Alessio Ishizaka, and Dan Nunan. 2020. Fake news, social media and marketing: a systematic review. *Journal of Business Research*, 124:329–341.

Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel Weld. 2008. Open information extraction from the web. *Commun. ACM*, 51:68–74.

Ernst Fehr and Urs Fischbacher. 2003. The nature of human altruism. *Nature*, 425:785–91.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Y. Bengio. 2014. Generative adversarial networks. *Advances in Neural Information Processing Systems*, 3.

Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. 2019. SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 845–854, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Jiaxian Guo, Sidi Lu, Han Cai, Weinan Zhang, Yong Yu, and Jun Wang. 2018. Long text generation via adversarial training with leaked information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Aditi Gupta, Hemank Lamba, Ponnurangam Kumaraguru, and Anupam Joshi. 2013. Faking sandy: Characterizing and identifying fake images on twitter during hurricane sandy. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13 Companion, page 729–736, New York, NY, USA. Association for Computing Machinery.

Stefan Helmstetter and Heiko Paulheim. 2018. Weakly supervised learning for fake news detection on twitter. In *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '18, page 274–277. IEEE Press.

Benjamin Horne and Sibel Adali. 2017. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 759–766.

Mahmoud Hossam, Trung Le, Michael Papasimeon, Viet Huynh, and Dinh Phung. 2021. Text generation with deep variational GAN. *CoRR*, abs/2104.13488.

Md. Sanzidul Islam, Sadia Sultana Sharmin Mousumi, Sheikh Abujar, and Syed Akhter Hossain. 2019. Sequence-to-sequence bangla sentence generation with lstm recurrent neural networks. *Procedia Computer Science*, 152:51–58. International Conference on Pervasive Computing Advances and Applications-PerCAA 2019.

Laurent Itti and Pierre Baldi. 2008. Bayesian surprise attracts human attention. *Vision research*, 49:1295–306.

Dongyeop Kang, Tushar Khot, Ashish Sabharwal, and Eduard H. Hovy. 2018. Adventure: Adversarial training for textual entailment with knowledge-guided examples. *CoRR*, abs/1805.04680.

Sanja Kelly, Mai Troung, Adrian Shahbaz, Madeline Earp, and Jessice White. 2017. Freedom on the net 2017. `https://freedomhouse.org/sites/default/files/2020-02/FOTN_2017_Final_compressed.pdf`.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*.

Angelika Kirilin and Michael Strube. 2018. Exploiting a speaker's credibility to detect fake news. In *Workshop on Data Science, Journalism and Media*.

Ramez Kouzy, Joseph Abi Jaoude, Afif Kraitem, Molly El Alam, Basil Karam, Elio Adib, Jabra Zarka, Cindy Traboulsi, Elie Akl, and Khalil Baddour. 2020. Coronavirus goes viral: Quantifying the covid-19 misinformation epidemic on twitter. *Cureus*, 12.

Sejeong Kwon, Meeyoung Cha, Kyomin Jung, Wei Chen, and Yajun Wang. 2013. Prominent features of rumor propagation in online social media. In *2013 IEEE 13th International Conference on Data Mining*, pages 1103–1108.

Pierre Lafon. 1980. Sur la variabilité de la fréquence des formes dans un corpus. *Mots*, 1:127–165.

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. 2017. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690.

Nayeon Lee, Belinda Z. Li, Sinong Wang, Pascale Fung, Hao Ma, Wen-tau Yih, and Madian Khabsa. 2021. On unifying misinformation detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5479–5485, Online. Association for Computational Linguistics.

Quanzhi Li, Qiong Zhang, and Luo Si. 2019. Rumor detection by exploiting user credibility information, attention and multi-task learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1173–1179, Florence, Italy. Association for Computational Linguistics.

Yaliang Li, Qi Li, Jing Gao, Lu Su, Bo Zhao, Wei Fan, and Jiawei Han. 2015. On the discovery of evolving truth. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, page 675–684, New York, NY, USA. Association for Computing Machinery.

Yunyao Li, Tyrone Grandison, Patricia Silveyra, Ali Douraghy, Xinyu Guan, Thomas Kieselbach, Chengkai Li, and Haiqi Zhang. 2020. Jennifer for COVID-19: An NLP-powered chatbot built for the people and by the people to combat misinformation. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.

Shuai Lin, Wentao Wang, Zichao Yang, Xiaodan Liang, Frank F. Xu, Eric Xing, and Zhiting Hu. 2020. Data-to-text generation with style imitation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1589–1598, Online. Association for Computational Linguistics.

Yunfei Long, Qin Lu, Rong Xiang, Minglei Li, and Chu-Ren Huang. 2017. Fake news detection through multi-perspective speaker profiles. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 252–256, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Sahil Loomba, Alexandre Figueiredo, Simon Piatek, Kristen de Graaf, and Heidi Larson. 2021. Measuring the impact of covid-19 vaccine misinformation on vaccination intent in the uk and usa. *Nature Human Behaviour*, 5:1–12.

Cristian Lumezanu, Nick Feamster, and Hans Klein. 2012. #bias: Measuring the tweeting behavior of propagandists. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 6, pages 210–217.

Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J. Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, page 3818–3824. AAAI Press.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2018. Rumor detection on Twitter with tree-structured recursive neural networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1980–1989, Melbourne, Australia. Association for Computational Linguistics.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2019. Detect rumors on twitter by promoting information campaigns with generative adversarial learning. In *The World Wide Web Conference*, WWW '19, page 3049–3055, New York, NY, USA. Association for Computing Machinery.

Xudong Mao, Qing Li, Haoran Xie, Raymond Y. K. Lau, and Zhen Wang. 2016. Multi-class generative adversarial networks with the L2 loss function. *CoRR*, abs/1611.04076.

Mohit Mayank, Shakshi Sharma, and Rajesh Sharma. 2021. Deap-faked: Knowledge graph based approach for fake news detection.

Philip M McCarthy. 2005. *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)*. Ph.D. thesis, The University of Memphis.

Philip M McCarthy and Scott Jarvis. 2010. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.

Liz McQuillan, Erin McAweeney, Alicia Bargar, and Alex Ruch. 2020. Cultural convergence: Insights into the behavior of misinformation networks on twitter. *CoRR*, abs/2007.03443.

Juan Carlos Medina Serrano, Orestis Papakyriakopoulos, and Simon Hegelich. 2020. NLP-based feature extraction for the detection of COVID-19 misinformation videos on YouTube. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.

Marcelo Mendoza, Barbara Poblete, and Carlos Castillo. 2010. Twitter under crisis: Can we trust what we rt? In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, page 71–79, New York, NY, USA. Association for Computing Machinery.

Lili Mou and Olga Vechtomova. 2020. Stylized text generation: Approaches and applications. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 19–22, Online. Association for Computational Linguistics.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:6859–6866.

Yingxue Pang, Jianxin Lin, Tao Qin, and Zhibo Chen. 2021. Image-to-image translation: Methods and applications. *IEEE Transactions on Multimedia*, PP:1–1.

Hao Peng, Ankur Parikh, Manaal Faruqui, Bhuwan Dhingra, and Dipanjan Das. 2019. Text generation with exemplar-based adaptive decoding. 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 2555–2565, Minneapolis, Minnesota. Association for Computational Linguistics.

Gordon Pennycook, Jonathon Mcphetres, Zhang Yunhao, and Jackson Lu. 2020. Fighting covid-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological Science*, 31:770–780.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. DeClarE: Debunking fake news and false claims using evidence-aware deep learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 22–32, Brussels, Belgium. Association for Computational Linguistics.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 231–240, Melbourne, Australia. Association for Computational Linguistics.

Alun Preece, Irena Spasic, Kieran Evans, David Rogers, William Webberley, Colin Roberts, and Martin Innes. 2017. Sentinel: A codesigned platform for semantic enrichment of social media streams. *IEEE Transactions on Computational Social Systems*, PP:1–14.

Alec Radford, Luke Metz, and Soumith Chintala. 2016. Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR*, abs/1511.06434.

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.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937, Copenhagen, Denmark. Association for Computational Linguistics.

Victoria Rubin, Niall Conroy, Yimin Chen, and Sarah Cornwell. 2016. Fake news or truth? using satirical cues to detect potentially misleading news. In *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*, pages 7–17, San Diego, California. Association for Computational Linguistics.

Sivasurya Santhanam. 2020. Context based text-generation using LSTM networks. *CoRR*, abs/2005.00048.

Leonard Savage. 1951. The theory of statistical decision. *JASA. Journal of the American Statistical Association*, 46.

Brett Schmidt. 2019. An exploration of text generation.

Shaden Shaar, Nikolay Babulkov, Giovanni Da San Martino, and Preslav Nakov. 2020. That is a known lie: Detecting previously fact-checked claims. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3607–3618, Online. Association for Computational Linguistics.

C. E. Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

D Sheridan, Danielle. 2019. Boris johnson to tackle anti-vaxx fake news on social media. https://www.telegraph.co.uk/news/2019/08/18/pm-say-social-media-firms-must-share-responsibility-rising-spread/. Accessed May 17, 2022.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2018. Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. *CoRR*, abs/1809.01286.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big Data*, 8:171–188.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017a. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19.

Kai Shu, Suhang Wang, and Huan Liu. 2017b. Exploiting tri-relationship for fake news detection. *ArXiv*, abs/1712.07709.

Kai Shu, Suhang Wang, and Huan Liu. 2019. Beyond news contents: The role of social context for fake news detection. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19, page 312–320, New York, NY, USA. Association for Computing Machinery.

Craig Silverman, Jeremy Singer-Vine, and Lam Vo, Thuy. 2017. In spite of the crackdown, fake news publishers are still earning money from major ad networks. https://www.buzzfeednews.com/article/craigsilverman/fake-news-real-ads#.iorLO7gqK. Accessed May 17, 2022.

Felix Simon, Philip Howard, N, and Rasmus Nielsen, Kleis. 2020. Types, sources, and claims of covid-19 misinformation. https://reutersinstitute.politics.ox.ac.uk/types-sources-and-claims-covid-19-misinformation. Accessed May 17, 2022.

Simranjeet Singh, Rajneesh Sharma, and Alan F Smeaton. 2020. Using gans to synthesise minimum training data for deepfake generation. *arXiv preprint arXiv:2011.05421*.

L Sydell, Laura. 2016. We tracked down a fake-news creator in the suburbs. here's what we learned. https://www.npr.org/sections/alltechconsidered/2016/11/23/503146770/npr-finds-the-head-of-a-covert-fake-news-operation-in-the-suburbs. Accessed May 17, 2022.

Eugenio Tacchini, Gabriele Ballarin, Marco Della Vedova, Stefano Moret, and Luca Alfaro. 2017. Some like it hoax: Automated fake news detection in social networks. Technical report, School of Engineering, University of California, Santa Cruz.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. 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 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Sebastian Tschiatschek, Adish Singla, Manuel Gomez Rodriguez, Arpit Merchant, and Andreas Krause. 2018. Fake news detection in social networks via crowd signals. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 517–524, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Svitlana Volkova, Kyle Shaffer, Jin Yea Jang, and Nathan Hodas. 2017. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on Twitter. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 647–653, Vancouver, Canada. Association for Computational Linguistics.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

Heng Wang, Zengchang Qin, and Tao Wan. 2017. Text generation based on generative adversarial nets with latent variable. *CoRR*, abs/1712.00170.

Lucy Wang, Arthi Ramachandran, and Augustin Chaintreau. 2021. Measuring click and share dynamics on social media: A reproducible and validated approach. *Proceedings of the International AAAI Conference on Web and Social Media*, 10(2):108–113.

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, Vancouver, Canada. Association for Computational Linguistics.

Claire Wardle and Eric Singerman. 2021. Too little, too late: Social media companies' failure to tackle vaccine misinformation poses a real threat. *BMJ*, 372:n26.

Wei Wei and Xiaojun Wan. 2017. Learning to identify ambiguous and misleading news headlines. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, page 4172–4178. AAAI Press.

You Wu, Pankaj Agarwal, Chengkai Li, Jun Yang, and Cong Yu. 2014. Toward computational fact-checking. *Proceedings of the VLDB Endowment*, 7:589–600.

Liang Xuyuan, Tian Lihua, and Li Chen. 2021. Tctg:a controllable text generation method using text to control text generation. In *2021 IEEE 6th International Conference on Signal and Image Processing (ICSIP)*, pages 1118–1122.

G York, Geoffrey. 2019. In worsening ebola outbreak, many congolese are shunning canadian-developed vaccine. https://www.theglobeandmail.com/world/article-in-worsening-ebola-outbreak-many-congolese-are-shunning-vaccine/. Accessed May 17, 2022.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. *CoRR*, abs/1905.12616.

Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas. 2017. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5908–5916.

Richard Zhang, Phillip Isola, and Alexei A. Efros. 2016. Colorful image colorization. In *Computer Vision – ECCV 2016*, pages 649–666, Cham. Springer International Publishing.

Yihao Zhao, Ruihai Wu, and Hao Dong. 2020. Unpaired image-to-image translation using adversarial consistency loss. *CoRR*, abs/2003.04858.

Fabiana Zollo, Alessandro Bessi, Michela Del Vicario, Antonio Scala, Guido Caldarelli, Louis M. Shekhtman, Shlomo Havlin, and Walter Quattrociocchi. 2015a. Debunking in a world of tribes. *CoRR*, abs/1510.04267.

Fabiana Zollo, Petra Kralj Novak, Michela Del Vicario, Alessandro Bessi, Igor Mozetič, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. 2015b. Emotional dynamics in the age of misinformation. *PLOS ONE*, 10(9):1–22.

Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018. Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys*, 51.