Understanding and Explicitly Measuring Linguistic and Stylistic Properties of Deception via Generation and Translation

Emily Saldanha
Data Science and Analytics Group
Pacific Northwest National Laboratory
emily.saldanha@pnnl.gov

Aparna Garimella
University of Michigan
(Now at Adobe Research)
garimell@adobe.com

Svitlana Volkova
Data Science and Analytics Group
Pacific Northwest National Laboratory
svitlana.volkova@pnnl.gov

Abstract

Massive digital disinformation is one of the main risks of modern society. Hundreds of models and linguistic analyses have been done to compare and contrast misleading and credible content online. However, most models do not remove the confounding factor of a topic or narrative when training, so the resulting models learn a clear topical separation for misleading versus credible content. We study the feasibility of using two strategies to disentangle the topic bias from the models to understand and explicitly measure linguistic and stylistic properties of content from misleading versus credible content. First, we develop conditional generative models to create news content that is characteristic of different credibility levels. We perform multi-dimensional evaluation of model performance on mimicking both the style and linguistic differences that distinguish news of different credibility using machine translation metrics and classification models. We show that even though generative models are able to imitate both the style and language of the original content, additional conditioning on both the news category and the topic leads to reduced performance. In a second approach, we perform deception style "transfer" by translating deceptive content into the style of credible content and vice versa. Extending earlier studies, we demonstrate that, when conditioned on a topic, deceptive content is shorter, less readable, more biased, and more subjective than credible content, and transferring the style from deceptive to credible content is more challenging than the opposite direction.

1 Introduction

As online social media usage continues to grow, it is becoming easier to access a much wider variety of news sources than ever before. Around two-thirds of U.S. adults get at least some of their news from sources on social media.[1] However, with the lack of traditional fact-checking and verification processes that accompany more standard news sources, this leads to a significant potential for the spread of false, misleading, and harmful information. The growing impact of such information in the online environment has lead to increased attention, awareness, and efforts to understand and combat its spread (Wardle and Derakhshan, 2017; Ireton and Posetti, 2018).

In social media, a great deal of attention has been dedicated to detect and measure the spread and impact of deceptive news (Lazer et al., 2018; Vosoughi et al., 2018). Many researchers (Pérez-Rosas and Mihalcea, 2015; Volkova et al., 2017; Rashkin et al., 2017; Wang, 2017; Baly et al., 2018) have analyzed linguistic differences to build models that classify types of deceptive news content. There are several limitations to the use of classification models to understand differences between deceptive and trustworthy content. Firstly, such models may learn to rely on the most prominent distinguishing lexical features between news categories but may not learn to model more subtle stylistic differences. Secondly, these models may learn context-dependent features that will evolve as the topics and news events change over time.

Our major contribution is to understand and explicitly measure linguistic and stylistic properties of content from misleading versus credible news sources while mitigating the topical bias. For that, we develop and rigorously evaluate two separate strategies: a *generation* approach to generate con-

[1] https://www.journalism.org/2018/09/10/news-use-across-social-media-platforms-2018/

Proceedings of The 13th International Conference on Natural Language Generation, pages 216–226,
Dublin, Ireland, 15-18 December, 2020.

tent of different credibility news source categories and a *translation* approach, which involves converting deceptive tweets to credible and vice versa while preserving the meaning. Our goal is to generate corresponding text in the style of credible and misleading news sources that will allow analysis of stylistic differences only, controlling for topical differences that would exist in the full corpus of real tweets.

First, for the text generation task, we aim to learn a generative language model that can produce news content characteristic of news sources of varied credibility in order to separate the linguistic and stylistic differences of these news sources. We aim to determine whether such models can reliably produce text that is characteristic of news sources of varied credibility, whether we can additionally control the topic of such generated text, and how we can best evaluate the performance of the generative models. We demonstrate that we can effectively generate text of different news source categories but that performance is reduced when additionally conditioning on topical indicators. Therefore, to perform linguistic analysis controlling for topic we turn to the translation task, for which we first create a parallel corpus and then take advantage of encoder-decoder architectures to train a transformation function to convert misleading to credible content and vice versa. Rather than aiming to modify the *information* content of the text, we aim to modify the stylistic properties while leaving the content as unchanged as possible.

2 Related Work

Efforts related to the detection of deception in written text has examined deceptive language in several domains such as fake news (Conroy et al., 2015), political speeches, online opinions about topics such as abortion or death penalty (Mihalcea and Strapparava, 2009; Newman et al., 2003). Most of the existing models for deception detection rely on linguistic features such as n-grams, language complexity, part-of-speech tags, and syntactic and semantic features (Mihalcea and Strapparava, 2009; Pérez-Rosas and Mihalcea, 2015; Yancheva and Rudzicz, 2013). Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2001) features were used to show that deceptive texts had fewer self-references, to avoid own self involved in the lies, and more words related to certainty (*always*, *all*, *truly*), possibly due to the need of the speaker to emphasize the "fake" truth (Mihalcea and Strapparava, 2009). Recent work in deception detection focused on developing predictive models to classify fake and verified news (Conroy et al., 2015; Rubin et al., 2016), assess information credibility (Wang, 2017), understand linguistic cues that distinguish fake and real news stories (Gravanis et al., 2019), and characterize the signatures of coordination (Alizadeh et al., 2020; Linvill and Warren, 2020). Recently, linguistically infused neural network architectures were developed for (a) classifying social media posts as credible or deceptive or as different types of deceptive news (Volkova et al., 2017), (b) factuality assessments of statements with different levels of credibility (Rashkin et al., 2017), and (c) classifying deceptive strategies and understanding intent behind deception (Volkova and Jang, 2018).

Text generation models leveraging deep learning architectures and neural language models have been applied to domains ranging from biographies (Lebret et al., 2016) to conversational text (Ghosh et al., 2017), and recently to generating rumors (Ma et al., 2019), fake news (Zellers et al., 2019) and stylometry (Schuster et al., 2020). Large-scale pre-trained language models, such as the generative pretrained transformer models, GPT-2 (Radford et al., 2019) and GPT-3 (Brown et al., 2020), have achieved excellent performance in free-form text generation, while approaches specifically for controllable text generation have been developed e.g., Plug and Play (Dathathri et al., 2020) and the Conditional Transformer Language (CTRL) model (Keskar et al., 2019).

Unlike any prior work on understanding the language of deception, we propose to disentangle the topic bias via translation and generation approaches to understand and measure stylistic and linguistic differences between misleading and credible news of different levels of credibility ranging from disinformation to credible news.

3 Data

While we would ideally like to identify deceptiveness at the level of *individual* news stories, due to the difficulty of this task and the high potential for misclassification, we instead focus on the classification of text at the news source level similar to (Lazer et al., 2018; Vosoughi et al., 2018). The news sources that we study are classified into the following categories based on publicly available lists of news sources annotated by experts.

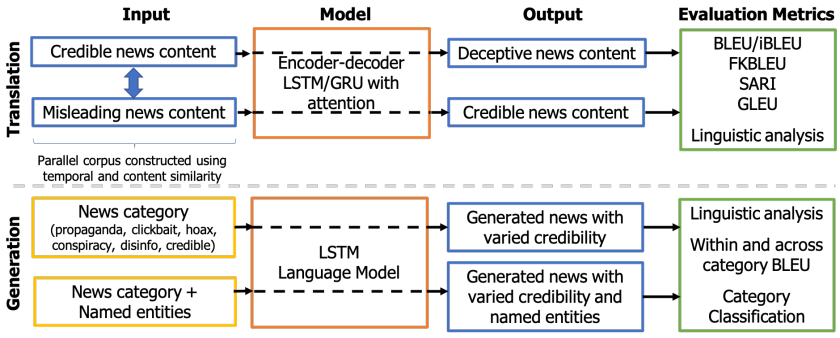


Figure 1: Conceptual framework for understanding and explicitly measuring linguistic and stylistic properties of deception via generation and translation.

- **Disinformation** sources provide partial, distorted, or false depictions of reality.
- **Conspiracy** sources tend to explain an event or practice by the coordinated actions of powerful people.
- **Propaganda** sources tend to persuade and manipulate public opinions and attitudes.
- **Hoax** sources seek to fool the gullible.
- **Clickbait** sources use attention-grabbing or vague headlines to drive engagement.
- **Credible** sources provide factual information with no intent to deceive the audience.

Further details about the annotation approach can be found in Volkova et al. (2017). Tweets were collected from the identified Twitter accounts from for the selected news sources during the 13-month period from January 2016 and January 2017 via the public Twitter API (Volkova et al., 2017). The distribution of tweets comprising 319K tweets from 231 deceptive news accounts, and 1M tweets from 340 credible news accounts.

Topical Indicators for Generation Task We aim to develop generation models that can generate text conditional on both news source type as well as specified topical content. As a proxy for content topic, we condition on different combinations of named entities mentioned in the tweets. We detect named entities in the tweets using the AllenNLP models[2] and annotate each tweet with a one-hot encoded indicator of the presence of each of the most common 250 named entities in the corpus.

Parallel Corpus Creation To support the translation task, we aligned tweet pairs from credible and deceptive news sources when they both were published on the same day (*temporal proximity*) and when the pair meets our threshold of *content similarity*. That is, when (a) all named entities and at least 70% (manually determined threshold) of the nouns in the deceptive tweet also occurs in the credible tweet, and (b) the (subject, verb, object) tuples obtained using SyntaxNet dependency parser in both deceptive and credible tweets match.

However, we do not want to have pairs with perfect similarities, as it would not serve our goal to study the differences between the style of credible and deceptive news sources. To obtain the final set of tweet pairs from the aligned set, we compute the text similarities for each pair using edit distance and cosine distances of Word2Vec, Doc2Vec, and TFIDF. After rigorous manual inspection we retain only those pairs with edit similarity that falls within the $(0.35, 0.7)$ range. This results in 105,365 aligned (credible, deceptive) tweet pairs. [3]

4 Approach

Generation Models To support our feasibility analysis, we leverage a three layer Long Short-Term Memory (LSTM) word-level language model for text generation. To create the conditional generative models, we add an embedding layer for the news source category that learns a dense representation for each of the six categories. The relevant embedding is appended to each token embedding in the input sequence, such that the model is tasked with predicting the next token given a modified version of the input token representation that is specific to the class of the input. When conditioning on named entity in addition to the news source, we add an additional component to the embedding by applying a dense layer to a one hot encoded representation of 250 most frequent named entities present in the tweet. For all models we use the

[2]https://demo.allennlp.org/named-entity-recognition

[3]The tweet data set, parallel corpus, and trained models will be made publicly available upon acceptance.

Adam optimizer, 256-dimensional hidden layers, a batch size of 256, and 200 epochs trained on a Tesla P40 GPU.

Translation Models Sequence-to-sequence neural models have significantly advanced the state-of-the-art in a variety of natural language processing tasks such as machine translation, speech recognition, and text summarization (Sutskever et al., 2014). We propose multiple variations of encoder-decoder models that learn a function that transforms tweets from deceptive news sources to the style of credible news sources, and the other way around. We run experiments with two commonly-used types of recurrent layers – LSTMs and Gated Recurrent Units (GRUs) – for the encoders and decoders. For both LSTMs and GRUs we used the Adam optimizer with a learning rate of 0.001, 256-dimensional embeddings, categorical cross-entropy loss, a batch size of 64, and 100 epochs with early stopping. Given the success of attention mechanisms for language tasks, we additionally experiment with multiplicative- (Luong et al., 2015) and self-attention (Vaswani et al., 2017) mechanisms.We conduct translations in both directions: (1) *forward*: credible to deceptive, and (2) *reverse*: deceptive to credible.

A conceptual schematic of both the generation and translation approaches is shown in Figure 1.

5 Generation Results

For the generation task, we aim to evaluate both the quality of the generated text and the similarity of the generated text to the desired conditional news category. The sign of good model performance is when the text similarity is higher between generated tweets of a specific category and real tweets of the same category than it is with real tweets of a different category.

We first evaluate the performance of the models that were conditioned on only the news source category. Examples of tweets of each category that were generated using these models can be found in Table 1. We next evaluate how the addition of topical conditioning affects the performance of the generative models. In particular, we are interested in how the additional conditioning affects the ability of the model to continue to generate tweets that are characteristic of the news source categories. Overall, we find that model performance on mimicking the differences between the news source types decreases due to an increasing complexity of the task with additional conditioning.

Generation Evaluation with BLEU In order to determine the textual similarity of tweets both within and across categories, we calculate the BLEU scores of both real and generated tweets from a given news source category with the same category and with different categories. Because we do not have one-to-one correspondence between real and generated text examples for the generation task, we calculate the BLEU score of each generated example against a sample of the full real corpus. We equivalently generate a BLEU score for a sample of real tweets against the real corpus to determine the typical n-gram similarity of real tweets to other real tweets. By comparing these two scores, we can determine how the similarity of generated samples compares with the expected similarity of real samples.

Figure 2 summarizes the BLEU score results for both real and generated content. We find that the BLEU score similarity to the target category is similar to what is observed for real content from the same category. However, the BLEU score similarity to content from categories other than the target category is significantly higher than what is observed for real cross-category content. For all categories we find that the BLEU score ratio of the generated content is greater than one, indicating that generated content is more similar to the target category than the other categories. However, in all cases the ratio is still significantly less than what is observed for the real content indicating that the models may be "hedging" and skewing their results towards similarity to the full corpus. We find that generated content that are characteristic of disinformation news sources come the closest to matching the BLEU-score ratio of real tweets.

For models which were conditioned on named entities, we can see that the raw BLEU scores are lower both within category and across category compared with the simpler model. This indicates that this model produces text that differs more from the real tweet corpus as a whole. Additionally, the BLEU score ratio has been reduced to below 1.0 for the hoax news category, indicating that these generated tweets are not more similar to real hoax tweets than real tweets of other categories. These results might be explained by the additional conditioning causing the generated distribution to differ from the real distribution. Because our goal is to

NEWS TYPE	GENERATED TWEETS
Clickbait	1) “ the congress is a dangerous guy . ” 2) this is the reason why the ‘ [OOV] ’ was the most popular isis group
Conspiracy	1) radioactive material stolen in dallas raises concern over # radioactive weapons 2) iran [OOV] $ [D] billion on finding a giant submarine [OOV]
Disinfo	1) # lavrov : we hope the # [OOV] is a long - term solution for russia ’ s foreign - [OOV] partnership 2) blast in # ankara after explosions heard in turkey : reports
Hoax	1) [D] times you ’ re all confident you don ’ t know what to see on your genes 2) girl loses her life in the world . . .
Propaganda	1) # erdogan : so much # isis will be forced to enter # syria [URL] # syriacrisis 2) # putin : russia ’ s # gas reserves soar at [OOV] since june [D][D][D][D]
Verified	1) punjab police arrest two suspects suspected of carrying explosives 2) # breaking : cnn projects [@] wins florida , democratic votes

Table 1: Example content generated for each news source type category using news source type conditional model.

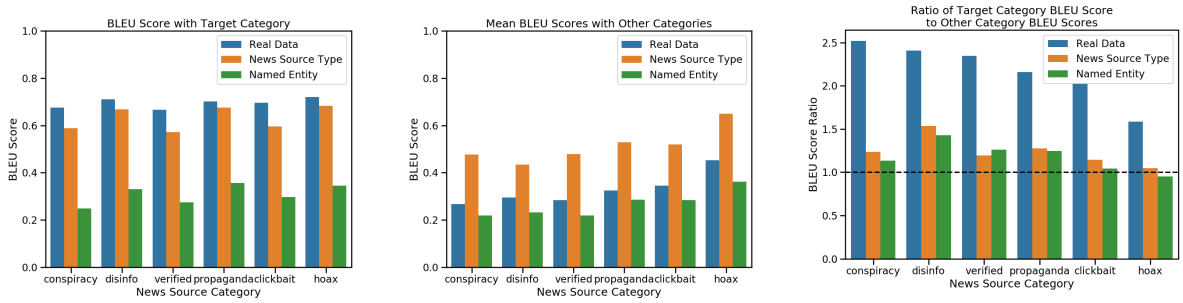


Figure 2: The BLEU scores of generated text (orange and green) calculated relative to real tweets of the target category (left) and real tweets of the other categories (middle), as well as the ratio of the BLEU score with the target category to the BLEU score with other categories (right). As a benchmark, the metrics for the generated text are compared with the BLEU scores of a sample of real tweets with other real tweets (blue).

generate text that stylistically matches the different news categories, even though we force the model to focus on topic areas that are not typical of that category, we aim to probe the stylistic similarity of real and generated text. Below we explore additional dimensions of evaluation to probe the stylistic quality of the conditional generation.

Generation Evaluation via Classification To probe the ability of the conditional model to mimic both content and stylistic differences among the news categories, we train models to predict news source category on both real and generated tweets. Note that the focus of this work is not to train the best deception detection model. Instead, it is to use the classification models to evaluate and contrast the performance of the generation models.

We evaluate each trained model on both real and generated tweets to determine how well the distinguishing patterns learned on each data set transfers to the other. We compare the performance across several different train and test splits. We train the classification model on the same tweets used to train the generative model and evaluate its performance on a validation set of real tweets and on the generated tweets. We also train the classification models on the validation set of real tweets and evaluate on the generated tweets. Finally, we train the classification model on the generated tweets and evaluate the performance on the real tweets from the generative training data, the real data from the validation set, and a set of held-out validation generated tweets.

To understand whether the generative models are successfully mimicking both the stylistic and linguistic differences of the news categories, we train models on several variations of the input text. First, we compare feed-forward models with bag-of-words (BoW) inputs to sequence-based LSTM models to compare the performance when given just topical word information versus stylistic phrasing information. Secondly, we compare using the true content words of the tweet with a processed version in which the content words are replaced with placeholders leaving only punctuation and other stylistic markers intact. If the classification models generalize well between the real and gener-

Data	Train Data	News Source Conditional F1 Real Train	F1 Real Val	F1 Gen	F1 Gen Val	Named Entity Conditional F1 Real Train	F1 Real Val	F1 Gen	F1 Gen Val
Full Text	Real Train	-	0.727	0.669	-	-	0.730	0.399	-
	Real Val	-	-	0.618	-	-	-	0.344	-
	Gen	0.293	0.291	-	0.813	0.264	0.264	-	0.673
BoW	Real Train	-	0.692	0.643	-	-	0.686	0.359	-
	Real Val	-	-	0.590	-	-	-	0.302	-
	Gen	0.304	0.301	-	0.878	0.246	0.237	-	0.790
No Content	Real Train	-	0.604	0.536	-	-	0.606	0.326	-
	Real Val	-	-	0.532	-	-	-	0.326	-
	Gen	0.324	0.324	-	0.636	0.252	0.253	-	0.525
No Content BoW	Real Train	-	0.492	0.442	-	-	0.490	0.275	-
	Real Val	-	-	0.411	-	-	-	0.271	-
	Gen	0.331	0.329	-	0.650	0.246	0.246	-	0.510

Table 2: Classification model performance for evaluation of the generation task, including performance for news source and named entity conditional generated tweets for different train and test sets used for the classification model and different data representations for the text. Performance on this classification task illustrates whether the generated texts have the same discriminative content and stylistic features as the real text.

ated text using just the stylistic information, we can infer that the generative models are able to model the tweet style in addition to the more salient content differences.

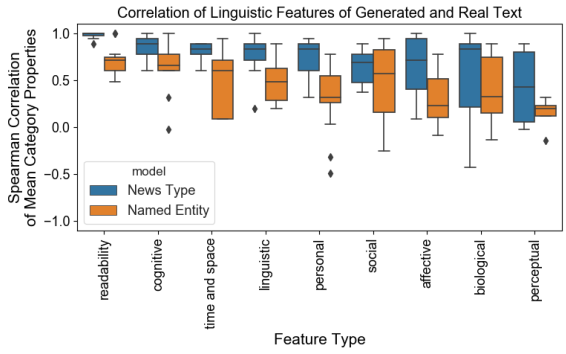
Table 2 summarizes the results of model evaluation via classification on the tweets. For the news source type conditional model, we find that the models trained on the real tweets are able to learn distinguishing features that generalize to the generated tweets, including both when topical information is included or excluded. This indicates that the generated content is replicating many of the features that characterized the differences among news source types. However, models trained on the generated tweets do not generalize well to the real text. This indicates that, while the generated tweets replicate many of the features that distinguish tweets of different credibility, they likely also include "cheats" that are not present in the real tweets.

We find that removing the content information reduces model performance. However, similar reduction in performance is observed when testing on both real and generated tweets when training on real data. This indicates that the generated tweets are retaining both content and style features that distinguish the different news categories. Interestingly, the performance of models trained on the generated tweets does not drop as much when the content is removed, indicating that for the generated tweets the structural characteristics are more distinct among categories.

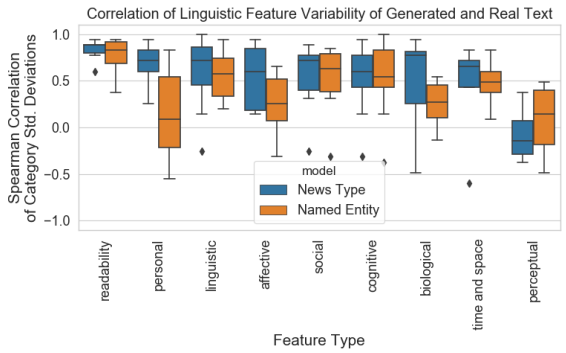
The classification performance on distinguishing the fine-grained credibility categories of the generated text sharply declines when conditioning on both news source type and the named entities. This indicates that the generated text is no longer replicating the same distinguishing features of the real text. Interestingly, this is also true for the classification models that leverage only stylistic information and not content information, indicating that the topical conditioning shifts the stylistic features of the text in addition to the content features.

Generation Evaluation via Linguistic Analysis As a final evaluation dimension, we compare the linguistic properties of generated text with real text using both readability measures and LIWC features. Because we are aiming to replicate the *differences* among text categories in the generated text, we evaluate the relative values of the linguistic measures among the news source categories. We first calculate the mean and standard deviation of each measure for each news source category of both generated and real text. Then we rank the different categories in terms of both their mean value and standard deviation and compare the observed rankings of the generated text with the observed ranking of the real text. This comparison is performed using the Spearman correlation. The comparison of the rankings of the mean value reveals whether the generated text follows the same relative linguistic patterns as the real text, while the comparison of the ranking of the standard deviation tells us

whether the generated text shows the same pattern of *variability* across categories as the real text.



(a) Average Linguistic Category Properties



(b) Linguistic Category Property Standard Deviations

Figure 3: Spearman correlation between the ranking of news categories of real tweets and generated tweets for the mean (a) and standard deviation (b) for different types of readability and LIWC measures. High correlation for the upper plot would mean that the generated text correctly replicated which categories has higher values of the linguistic features than others, while high correlation in the lower plots would mean that the generated text correctly replicated which categories had higher variability in linguistic properties than others.

Figure 3 presents the comparison of the relative linguistic feature values between the real and generated tweets. The generated text replicated the observed patterns in the readability statistics among categories very well, with an average Spearman correlation of the mean readability metrics of 0.98 and the average correlation of the standard deviation of these metrics of 0.84.

The average Spearman correlation of the mean LIWC features values among categories was 0.71 and the average correlation of the standard deviation of the features was 0.53. We find that on average all types of LIWC features have positive correlation, but that the cognitive type features, e.g. *causality and certainty, time and space* features, e.g. *motion and directions*, and the linguistic type features, e.g. *pronoun usage and negation*, have the strongest agreement with the real text. The LIWC features with low correlation between the generated and real tweets tend to be those features which leverage keywords that are relatively rare in the true data.

In comparison with the simpler conditional models, we find a lower average Spearman correlation for both readability and LIWC measures for the named entity conditional model. For the readability measures, the correlation of the mean values was 0.70 and the correlation of the standard deviations was 0.76. For the LIWC features, the correlation of the mean values was 0.41 and the correlation of the standard deviations was 0.35.

Generation Results Summary We have leveraged three different evaluation approaches to probe the ability of the generation models to mimic both the language and the style of different news categories. These metrics have consistently shown that the news source category conditional models are able to successfully replicate distinguishing features of the different categories in terms of both content and style. However, this ability, in terms of both style and content, is significantly reduced when conditioning on additional topical information in the form of named entities. Because we do not achieve sufficient performance of the topic conditional models, further development is needed before these models can be leveraged to control for topic bias in analysis of the stylistic differences between credible and deceptive news sources.

6 Translation Results

To evaluate the translation task, we rely on a set of metrics used to evaluate machine translation and text generation systems: BLEU (Papineni et al., 2002), *i*BLEU (Sun and Zhou, 2012), FK-BLEU (Xu et al., 2016), SARI (Xu et al., 2016), and GLEU (Wu et al., 2016). We present performance results of translations between deceptive and credible aligned tweets in Table 3. All encoder-decoder models that learn transformations from credible into deceptive (*forward* translation) are more accurate than those that learn translations from deceptive into credible (*reverse* translation). GRUs perform significantly better than LSTMs, while GRUs perform even more accurately when augmented with the attention mechanism. The addition of self-attention results in the highest trans-

	Credible→Deceptive				Deceptive→Credible			
Metric	LSTM	GRU	GRU+MA	SA	LSTM	GRU	GRU+MA	SA
BLEU	17.79	19.97	20.42	**37.19**	16.97	18.89	20.36	**33.91**
*i*BLEU	14.43	16.35	17.21	**31.31**	13.65	15.51	16.92	**28.07**
FKBLEU	8.16	8.23	9.61	**16.31**	7.87	5.41	9.54	**10.46**
SARI	38.53	39.52	39.21	**47.87**	37.85	38.46	38.54	**45.53**
GLEU	9.51	11.50	11.76	**28.07**	6.73	7.94	9.65	**18.83**

Table 3: Encoder-decoder model performance for translations from deceptive to credible content and vice versa. MA refers to models with multiplicative attention and SA refers to models with self-attention.

Deceptive Source	Credible Target	Model Output	SARI
russian experts to fly to turkey to investigate ambassadors murder	russian ambassador to turkey attacked at photo exhibition	russian ambassador to turkey shot dead at photo exhibition in ankara	0.64
north korea claims to have successfully carried out its fifth nuclear test	north korea is believed to have conducted a fifth nuclear test	north korea conducts fifth and largest nuclear test	0.41

Table 4: Examples of translations from misleading news sources to credible using our best model with the self-attention.

lation performance in both the forward and the reverse directions. To qualitatively demonstrate model performance, we present example outputs including good and bad translations from our best model in Table 4.

6.1 Topic-Controlled Linguistic and Stylistic Analysis

We leverage our learned translation models to examine the stylistic properties of credible versus misleading news sources independent of the topic. We perform pairwise comparative linguistic analysis of real tweets from credible news sources in comparison with their corresponding translated deceptive counterparts using readability, biased and subjective language measures. We also compare stylistic features including the average number of edit operations (insertions, deletions, and substitutions) required to transform a tweet from a credible news source into one from a deceptive one, and vice versa, and differences in tweet length. We report the results for each measure in Table 5.

We find that tweets from credible news sources are longer, on average, with an average length of 73 characters compared to 68 for misleading content. Consistently, we find that an average of 6 character deletions are needed to go from credible to deceptive and 6 insertions to transform vice versa.

To compare *readability* of content from credible and misleading news sources conditioned on the topic of the post, we apply several widely-used readability measures: Automated Readability Index (ARI) (Senter and Smith, 1967), Flesch Reading Ease (FRE) (Farr et al., 1951), Flesch-Kincaid Grade Level (FKGL) (Kincaid et al., 1975), the

Measure	Credible	Deceptive
Readability		
ARI	9.309	**9.970**↑
FK Grade Level	10.004	**10.822**↑
Complex words	2.515	**2.535**↑
Flesch Reading Ease	**44.893**↑	36.412
Syllables	**21.053**↑	19.526
Subjective Language		
Strongly positive	0.024	**0.025**↑
Strongly negative	0.023	**0.025**↑
Weakly positive	0.026	**0.033**↑
Weakly negative	0.065	**0.066**↑
Biased Language		
Assertive verbs	**0.017**↑	0.013
Report verbs	**0.042**↑	0.038
Implicative verbs	0.009	**0.010**↑
Stylistic Properties		
No. of Characters	**73.898**↑	68.209
Insertions	10.540	**16.2276**↑
Substitutions	32.044	32.0482

Table 5: Linguistic differences between parallel content from credible and deceptive sources controlling for topic bias. Stat. sign. differences are shown in bold (Mann–Whitney U test $p < 0.005$).

number of syllables, and the number of complex words. A given tweet is more readable when its ARI and FKGL are low, but FRE is high. On an average, narratives from credible news sources are more readable compared to narratives from misleading news sources – i.e., they have lower ARI and FKG, higher FRE scores, have more syllables, and fewer complex words.

We use publicly available *subjectivity* lexicons (Riloff and Wiebe, 2003; Liu et al., 2005) to annotate strongly negative, weakly negative, weakly positive, and strongly positive terms in the tweets. We contrast the average fraction of terms in each tweet that belong to one of the four groups in tweets from credible and deceptive news sources. In line with earlier work, we confirm that content from deceptive sources has more subjective terms, even when conditioned on topic.

We compute the average fraction of terms in each tweet that fall under one of the following verb types: assertive verbs (Hooper, 1974) that bring emphasis to a sentence (*point out*, *claim*), implicative verbs (Karttunen, 1971) whose factuality depends upon a condition (*avoid*, *hesitate*), and report verbs (Recasens et al., 2013) that indicate that discourse is being quoted or paraphrased (*admit*, *criticize*). We found that controlling for topic, credible content has a higher fraction of assertive and report verbs, while implicative verbs occur more often in deceptive tweets.

If we compare these results to existing results which directly compare tweets from credible and deceptive sources without controlling for topic, we find that some conclusions are confirmed by this topic-controlled analysis while others show a discrepancy. Consistent with (Volkova et al., 2017) and (Rashkin et al., 2017), we find that deceptive news sources use more subjective and biased language than credible sources. However, in contrast with the earlier studies, we find that after controlling for topic credible news sources are more likely to use assertive and report verbs.

7 Conclusions and Future Work

We have presented a novel understanding of linguistic and stylistic properties deception using translation and generation approaches designed to disentangle from the topic bias.

We have demonstrated progress towards developing generative models for this purpose, with our generative models being able to imitate both the style and the content of the real tweets. We have evaluated our generative models using BLEU scores, linguistic analysis, and classification. However, with additional topical conditioning dimensions, we find significantly reduced performance on maintaining the observed stylistic and content differences. In order to improve the performance of these generative models, future work includes the application of more advanced controllable text generation model transformer-based and other recently emerged neural architectures (Brown et al., 2020; Prabhumoye et al., 2020; Keskar et al., 2019; Dathathri et al., 2020; Radford et al., 2019). We will focus on improving the ability to perform multi-dimensional conditioning to enable the desired topic-controlled misleading versus credible news source style analysis.

We have shown that translating from misleading to credible content is more difficult than the opposite direction. This may be because deceptive content has a higher level of stylistic and lexical variation making it more difficult for the models correctly anticipate its style, while verified content has a more factual style. Our translation results clearly demonstrate that, when conditioned on a topic, content from deceptive news sources is shorter, less readable, more biased, and more subjective than content from deceptive news sources. Leveraging these models to allow topic-agnostic comparison of the style of deceptive and credible news sources, we demonstrate several key stylistic differences.

References

Meysam Alizadeh, Jacob N Shapiro, Cody Buntain, and Joshua A Tucker. 2020. Content-based features predict social media influence operations. *Science advances*, 6(30):eabb5824.

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 Empirical Methods in Natural Language Processing*, pages 3528–3539.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Nadia K Conroy, Victoria L Rubin, and Yimin Chen. 2015. Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology*, 52(1):1–4.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: a simple approach to controlled text generation. *International Conference on Learning Representations*.

James N Farr, James J Jenkins, and Donald G Paterson. 1951. Simplification of Flesch Reading Ease Formula. *Journal of Applied Psychology*, 35(5):333.

Sayan Ghosh, Mathieu Chollet, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer. 2017. Affect-LM: A neural language model for customizable affective text generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 634–642, Vancouver, Canada. Association for Computational Linguistics.

Georgios Gravanis, Athena Vakali, Konstantinos Diamantaras, and Panagiotis Karadais. 2019. Behind the cues: A benchmarking study for fake news detection. *Expert Systems with Applications*, 128:201–213.

Joan B Hooper. 1974. *On assertive predicates*. Indiana University Linguistics Club.

Cherilyn Ireton and Julie Posetti. 2018. *Journalism, fake news & disinformation: handbook for journalism education and training*. UNESCO Publishing.

Lauri Karttunen. 1971. Implicative verbs. *Language*, pages 340–358.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. *Science*, 359(6380):1094–1096.

Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1203–1213, Austin, Texas. Association for Computational Linguistics.

Darren L Linvill and Patrick L Warren. 2020. Troll factories: Manufacturing specialized disinformation on twitter. *Political Communication*, pages 1–21.

Bing Liu, Minqing Hu, and Junsheng Cheng. 2005. Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web*, pages 342–351.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. 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*, pages 3049–3055.

Rada Mihalcea and Carlo Strapparava. 2009. The lie detector: Explorations in the automatic recognition of deceptive language. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 309–312. Association for Computational Linguistics.

Matthew L Newman, James W Pennebaker, Diane S Berry, and Jane M Richards. 2003. Lying words: Predicting deception from linguistic styles. *Personality and social psychology bulletin*, 29(5):665–675.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Verónica Pérez-Rosas and Rada Mihalcea. 2015. Experiments in open domain deception detection. In *Proceedings of Empirical Methods in Natural Language Processing*, pages 1120–1125.

Shrimai Prabhumoye, Alan W Black, and Ruslan Salakhutdinov. 2020. Exploring controllable text generation techniques. *arXiv preprint arXiv:2005.01822*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 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 Empirical Methods in Natural Language Processing*, pages 2931–2937.

Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1650–1659, Sofia, Bulgaria. Association for Computational Linguistics.

Ellen Riloff and Janyce Wiebe. 2003. Learning extraction patterns for subjective expressions. In *Proceedings of Empirical Methods in Natural Language Processing*.

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.

Tal Schuster, Roei Schuster, Darsh J Shah, and Regina Barzilay. 2020. The limitations of stylometry for detecting machine-generated fake news. *Computational Linguistics*, pages 1–12.

RJ Senter and Edgar A Smith. 1967. Automated readability index. Technical report.

Hong Sun and Ming Zhou. 2012. Joint learning of a dual SMT system for paraphrase generation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 38–42, Jeju Island, Korea. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Svitlana Volkova and Jin Yea Jang. 2018. Misleading or falsification: Inferring deceptive strategies and types in online news and social media. In *Companion Proceedings of the The Web Conference 2018*, pages 575–583.

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.

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 Hossein Derakhshan. 2017. Information disorder: Toward an interdisciplinary framework for research and policy making. *Council of Europe report*, 27.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Maria Yancheva and Frank Rudzicz. 2013. Automatic detection of deception in child-produced speech using syntactic complexity features. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 944–953.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In *Advances in Neural Information Processing Systems*, pages 9054–9065.