Towards Trustworthy Deception Detection: Benchmarking Model Robustness across Domains, Modalities, and Languages

Maria Glenski, Ellyn Ayton, Robin Cosbey, Dustin Arendt, and Svitlana Volkova
Pacific Northwest National Laboratory
Richland, WA, USA
first.last@pnnl.gov

Abstract

Evaluating model robustness is critical when developing trustworthy models not only to gain deeper understanding of model behavior, strengths, and weaknesses, but also to develop future models that are generalizable and robust across expected environments a model may encounter in deployment. In this paper we present a framework for measuring model robustness for an important but difficult text classification task – deceptive news detection. We evaluate model robustness to out-of-domain data, modality-specific features, and languages other than English.

Our investigation focuses on three type of models: LSTM models trained on multiple datasets (Cross-Domain), several fusion LSTM models trained with images and text and evaluated with three state-of-the-art embeddings, BERT ELMo, and GloVe (Cross-Modality), and character-level CNN models trained on multiple languages (Cross-Language). Our analyses reveal a significant drop in performance when testing neural models on out-of-domain data and non-English languages that may be mitigated using diverse training data. We find that with additional image content as input, ELMo embeddings yield significantly fewer errors compared to BERT or GLoVe. Most importantly, this work not only carefully analyzes deception model robustness but also provides a framework of these analyses that can be applied to new models or extended datasets in the future.

1 Introduction

Detection of deceptive content online is an extremely important but challenging task and there have been significant efforts to apply machine learning and deep learning to solve this problem (Rubin et al., 2016; Mitra et al., 2017; Wang, 2017; Karadzhov et al., 2017; Volkova et al., 2017; Shu et al., 2017; Rashkin et al., 2017). A plethora of models exist that rely on a range of data mined from various social platforms – Facebook, Twitter, Reddit – and modalities – text, images, or both. These existing models rely on different text features (*e.g.,* linguistic structure, lexical, psycholinguistic features, biased and subjective language), image features, and user interaction features (*e.g.,* engagement, network structure, temporal patterns). However, current literature lacks details for how these models transfer, *e.g.,* to out-of-domain data, across platforms, or across languages.

Moreover, there is a gap in understanding the underlying behavior of the decision-making processes behind model output. It is not clear why models for deception detection (particularly neural or deep learning based models) are making certain predictions (*e.g., why* one item is predicted as deceptive versus not). Evaluation of model performance with one metric such as accuracy or F1 score is not enough. For such complex prediction tasks like deception detection for which humans often disagree (Karduni et al., 2018; Karduni et al., 2019; Ott et al., 2011; Harris, 2012), we need more rigorous evaluation of neural model behavior and a cohesive system for comparing model results across bodies of research.

Evaluations need to explicitly measure the extent to which model performance is affected when new data is supplied, *e.g.,* data with a different topic distribution or how well-performing models on English

Proceedings of the 3rd International Workshop on Rumours and Deception in Social Media (RDSM), pages 1–13
Barcelona, Spain (Online), December 12, 2020.

data will perform on non-English inputs. There is also a critical need for evaluations that highlight when a model is correct, which examples have the ability to explain why, and, most importantly, conclude why a user should trust a given model in an interpretable manner. Arguments in favor of the above requirements to model performance are well-aligned with recent work on machine learning interpretability, trust, fairness, accountability, and reliability (Lipton, 2018; Doshi-Velez and Kim, 2017; Hohman et al., 2018). Another key argument in favor of rigorous evaluation is the lack of benchmark datasets and the need for reproducibility. Often, researchers do not make models or datasets publicly available and the details to reconstruct their models and experimental setup are not sufficient, which prevents other researchers from performing rigorous comparisons. To reproduce our key findings and experiments, we will make our framework available through interactive Jupyter notebooks available at publication.

The focus of the current work and our main contribution is an *extensive evaluation of neural model robustness across frequently encountered factors in real-world applications of digital deception models* such as new domains, languages, modalities, and upgrades to new state-of-the-art text embeddings.

2 Related Work

Deep learning systems have been adopted in many areas from medicine to autonomous driving (Ahmad et al., 2018; Claybrook and Kildare, 2018) and as these algorithms are incorporated, the need for explainable and transparent models becomes more urgent. One approach researchers use to overcome the inherent ambiguity of these black-box methods is to develop additional models to learn and explain the decisions of existing models, analyze when these models fail, and introduce a human-in-the-loop component to improve performance (Ribeiro et al., 2016; Murdoch et al., 2019; Poursabzi-Sangdeh et al., 2018). Another way to tackle this challenge is to design models with a goal of interpretability in place when development starts (Ridgeway et al., 1998; Rudin, 2018; Gilpin et al., 2018; Lahav et al., 2018; Hooker et al., 2019). Model performance beyond accuracy, precision, and recall measures are essential in order to build trustworthy and reliable models tasked with making decisions capable of significantly affecting the end users (Hohman et al., 2018; Dodge et al., 2019b; Hohman et al., 2019).

In this study, we do not focus on developing new visualization techniques to explain deception detection models using intrinsic or post-hoc explanations (Yang et al., ; Shu et al., 2019a; Reis et al., 2019; Wallace et al., 2019) or measure their interpretability (Mohseni et al., 2019). Instead, we thoroughly evaluate model performance using error analysis under several real-world conditions e.g., across domains and input types to understand reliability and the underlying behavior of decision making processes that will in turn increase model explainability and interpretability.

Deception Detection Detecting suspicious or deceptive news online is a well explored area of study. Current research efforts focus on broadly classifying between suspicious and trustworthy content (Volkova et al., 2017; Shu et al., 2020) to more specific distinctions such as between propaganda, hoax, satire, and trustworthy news (Rashkin et al., 2017), to examining the behavior of malicious users and bots (Glenski and Weninger, 2018; Kumar et al., 2017; Kumar et al., 2018), or analyzing misinformation and rumor spread patterns over time (Kwon et al., 2017; Vosoughi et al., 2018). Methods for classification vary from random forests to deep neural networks and the addition of enriched features such as images, temporal and structural attributes, and linguistic features have been shown to boost model performance over relying on textual characteristics alone (Wang, 2017; Qazvinian et al., 2011; Kwon et al., 2013).

With the high impact of digital deception on offline, real-world events, effective detection models are a critical concern. However, interpretable evaluations of model performance beyond traditional precision, recall, or F score measures are essential to building trustworthy and reliable systems when deep learning or machine learning models are tasked with making decisions capable of significantly affecting end users (Dodge et al., 2019b; Volkova et al., 2019; Hohman et al., 2019; Hohman et al., 2018).

Many studies have relied on Recurrent Neural Networks (RNNs) or variations, particularly Long Short Term Memory (LSTM) layers, for deception detection tasks (Ma et al., 2016; Chen et al., 2018; Rath et al., 2017; Zubiaga et al., 2018; Zhang et al., 2019). Others have used Convolutional Neural Networks (CNN) (Ajao et al.,), or variations of LSTM architectures such as including attention mechanisms (Guo et al., ; Li et al., 2019) which are typically dependent on specific tasks or parameter tuning of state-of-the-art deception detection models. For the purposes of consistency across our experiments, we rely on standard LSTM models that underpin much of the recent work rather than tuning each architecture by task and data or comparing the wider range of recent state-of-the-art architectures. This enables us to make more accurate comparisons of the factors related to robustness that we vary in our experiments. Developing novel models or comparing all state-of-the-art architectures is beyond the scope of this paper.

Evaluating Model Robustness Measuring robustness of a model goes beyond fortifying against intentional manipulations. Corrupted data inputs are possible in real-world scenarios without harboring the intent to fool a system and models should behave as expected in such cases, *e.g.,* simple image transformations, missing or misspelled text. Several key studies have endeavored to develop methods to evaluate model robustness (Zheng et al., 2016; Hein and Andriushchenko, 2017; Liu et al., 2018; Dodge et al., 2019a). In this study we focus on evaluating model robustness to out-of-domain data, multilingual data, and multimodal inputs combined with conceptually different text embedding techniques to gain deeper insights into how much "learning and understanding" neural network models actually have. Different studies have separately examined the relationships between extracted features across modalities, across languages, and across domains (Zhou et al., 2020; Capuozzo et al., 2020; Zotova et al., 2020). To the best of our knowledge, this extensive evaluation of deception detection model robustness across multiple tasks has not been presented before in one unified work.

3 Methodology

In this section we discuss the state-of-the-art neural models for deception detection used in this paper and describe our approach for evaluating model robustness across domains, modalities, and languages, a summary of which is presented in Table 1.

3.1 Neural Models for Deception Detection

After an extensive analysis of published neural deception classification models, we adopt similar architectures that consist of a commonly used RNN architecture, LSTM layers, and rely on combinations of text, linguistic cue (LC), and image vectors for our experiments. Lexical vectors and linguistic cues include those often used for classification, notably encoding for biased and otherwise subjective language (Rashkin et al., 2017; Shu et al., 2019b). We employ LIWC (Pennebaker et al., 2001) and several lexical dictionaries (assertive verbs, hedges, factives, implicatives, etc. (Recasens et al., 2013)) to construct frequency vectors for model input. Image vectors are representations extracted from the last layer of the state-of-the-art ResNet architecture (He et al., 2016). We focus our analysis on both models with single modalities *e.g.,* text and multimodal models *e.g.,* text and image, as well as classification tasks over a varied number of classes. We create 80/20 splits in each dataset for train and test sets.

We implement the binary classification model over trustworthy and deceptive samples using text and lexical input features to evaluate the efficacy and robustness of testing on out-of-domain data. We additionally implement the binary (trustworthy versus deceptive), the 3-way (trustworthy, propaganda, disinformation), and the 4-way (clickbait, satire, hoax, conspiracy) classification models with the enhancement of image input features for our evaluation of multimodal models. Our final model to assess the robustness of these deep learning approaches is the 3-way classification of clickbait, conspiracy, and

Model	Platform(s)	Input	Text Embeddings	Detection Task (Classes)	Robustness Task
M_1	Twitter Reddit	Text + LC	BERT	trustworthy, deceptive	Cross-Domain
M_2	Twitter	Text + Image + LC	BERT, GLoVe, ELMo	trustworthy, deceptive	Cross-Modality
M_3	Twitter	Text + Image + LC	BERT, GLoVe, ELMo	trustworthy, propaganda, disinformation	Cross-Modality
M_4	Twitter	Text + Image + LC	BERT, GLoVe, ELMo	clickbait, satire, hoax, conspiracy	Cross-Modality
M_5	Twitter	Text	Character	clickbait, conspiracy, propaganda	Cross-Language

Table 1: Overview of the deception detection models (M_1, M_2, M_3, M_4, and M_5) used for each task.

	Train		**Test**	
Data	*Trustworthy*	*Deceptive*	*Trustworthy*	*Deceptive*
Twitter	14.9k	28.0k	3.7k	7.2k
Reddit	21.6k	21.5k	5.4k	5.5k
Twitter + Reddit	36.6k	49.9k	–	–

Table 2: Distributions across classes within train and test data used in the cross-domain analyses (M_1).

propaganda news sources across multiple languages. We employ similar neural network architectures for all models, which consists of a pre-trained text embedding layer followed by an LSTM layer and a single dense layer. The output from this sub-network is concatenated to the image vectors which have been transformed by two dense layers. In the architectures of our multi-modal models, our joint text and image representations are concatenated to the lexical cues vector as input to the final two fully connected layers in the model. All layers use dropout.[1]

In contrast to the Cross-Domain and Cross-Modality models, the multilingual Cross-Language models do not incorporate linguistic cues or other lexical features because of the inconsistency across multiple languages – some of the linguistic features and lexicons are not available for all languages (*e.g.,* bias and subjective language dictionaries). The multilingual model (M_5) architecture is constructed by the pre-trained character-level text embedding layer followed by two convolution layers, then a max pooling layer, and finally two dense, fully connected layers.

3.2 Datasets

We use a previously released, public list of news sources[2] annotated along a spectrum of deceptive – clickbait, hoax, satire, conspiracy, propaganda, or disinformation – and verified news sources who typically spread factual content (which we denote as “trustworthy” in our experiments). These annotated news sources focused on activity in 2016, therefore, we collected the following datasets for the same 12 month period of activity. Note, there are limitations when annotations are done on the sources level, however, similar to related work (Vosoughi et al., 2018; Lazer et al., 2018; Glenski et al., 2018) we advocate focusing on the news sources rather than individual stories because we view the definitive element of deception to be the intent and the tactics of the news source.

In our ***Cross-Domain*** analyses, we consider two domains – Twitter and Reddit. The Twitter component of the M_1 dataset is comprised of English retweets from official twitter accounts for news media described above, containing only text-based posts. The Reddit component comprises all top-level comments in response to posts on Reddit in 2016 where the post contained a link to a web domain associated with one of the news accounts of interest. We create a binary classification dataset, analogous to the Twitter dataset used for M_2, that collapses news source annotations to either trustworthy or deceptive and down-samples the Reddit comments to have approximately the same volume of content ($N = 54k$). This allows the joint *Twitter + Reddit* dataset to be balanced between the two domains for cross-domain analyses, as shown in Table 2 where we present the class distributions for the M_1 dataset.

Our multimodal Twitter data, for the ***Cross-Modality*** analyses, is comprised of the same collection of English retweets from annotated news media’s official accounts used to build the Twitter component of the M_1 dataset, filtered to those tweets that include images. Each retweet comprises a unique image and body of text. We create datasets for three classification tasks: (1) propaganda, disinformation, or trustworthy (M_3; $N = 54.4k$), (2) clickbait, conspiracy, hoax, or satire (M_4; $N = 2.5k$), and (3) a binary classification dataset (M_2; $N = 54.4k$) of trustworthy versus deceptive tweets in the dataset used for M_3 by collapsing the sub-classes of propaganda and disinformation into a single deceptive class. The

[1] For each model, we perform a random hyper-parameter tuning search independently consisting of ten different configurations of model hyper-parameters. The batch size, the number of training epochs, drop out rate, and the recurrent layer dimensionality as well as the optimizer and learning rate are among the tunable parameters. Every model had a unique set of final hyper-parameters, however, the most common configurations consisted of the Adam optimizer with a learning rate of 0.0001, a drop out rate between 0.2 and 0.25, and 10 epochs of training.

[2] We use the following publicly available news source annotations for *deceptive news sources*: `https://www.cs.jhu.edu/~svitlana/data/SuspiciousNewsAccountList.tsv` and *trustworthy news sources*: `https://www.cs.jhu.edu/~svitlana/data/VerifiedNewsAccountList.tsv`

breakdown of these three tasks is presented in Table 1.

In contrast to the previous, English-only datasets, our multilingual Twitter dataset for the ***Cross-Language*** analyses comprises 7,316 tweets across five languages (English, French, German, Russian, and Spanish) from a similar collection of tweets as described above for English Twitter without the restriction to only English-text. Because of an over-representation of English-tweets in the original collection, we sample 1,500 examples for each language evenly distributed across the three classes, *i.e.,* 500 clickbait samples, 500 conspiracy samples, and 500 propaganda samples. For the least represented language, Russian, we use all available tweets: 478 clickbait samples, 338 conspiracy samples, and 500 propaganda samples. For all languages, we allocate 80% to train and 20% to test.

3.3 Evaluating Neural Model Robustness

We investigate the robustness of digital deception models when evaluated across three dimensions: domains, modalities and languages, seeking to answer three key research questions. Table 1 outlines the model architectures, inputs, datasets, and prediction tasks for all robustness analyses.

In our ***Cross-Domain*** analyses, we evaluate model robustness across two popular social platforms: Twitter and Reddit. We compare model performance on a binary classification task (M_1 in Table 1) using the binary English Twitter and Reddit datasets summarized in Table 2. We train three LSTM models using textual and lexical input features – (1) trained on Twitter content only, (2) trained on Reddit content only, and (3) trained on both the Twitter and Reddit content. All three models rely on pre-trained BERT word embeddings that are fine-tuned during training and use the same neural architecture and hyper-parameters described in section 3.1. To account for inherent platform inconsistencies between Twitter and Reddit, we limit the input text length to 100 words and use a common vocabulary between the three model setups. We evaluate each model on two held out test sets composed of held out examples from Twitter and Reddit allowing us to test performance within the same domain (*e.g.,* train and test on Twitter data) and on out-of-domain data (*e.g.,* train on Twitter and test on Reddit).

In our ***Cross-Modality*** analyses, we evaluate how multiple modalities can influence the predictions of a model. We train three LSTM classifiers to predict falsified news at different levels of granularity. The first model (M_2) is trained to distinguish between trustworthy and deceptive news using text, lexical cues, and extracted image features. The second model (M_3) uses more fine-grained labels of digital deception to further separate deceptive news into propaganda and disinformation. Finally, (M_4) is a 4-way classification task to predict tweets from four types of deceptive news types: clickbait, conspiracy, hoax, and satire. Each classifier is trained three separate times using a different text embedding strategy that allows for fine-tuning during training: BERT, ELMo, or GloVe. We use the Tensorflow Object Detection API (Huang et al., 2017) to extract the primary COCO[3] (Common Objects in COntext) objects from the images of the tweets. In this analysis, we also seek to answer, in particular, the question: *Which errors do text embeddings or detected objects contribute to in the multimodal setup?*

In our ***Cross-Language*** analyses, we evaluate changes in model performance when a convolutional neural network (M_5) developed to distinguish fine-grained differences in digital deception (*clickbait* versus *conspiracy* versus *propaganda*) is trained and evaluated across multiple languages. We examine the impacts on performance when the model is trained in the context of multiple languages (*i.e.,* a single model trained using an aggregated dataset of English, Russian, German, Spanish, and French samples) and evaluated on a single language as well as when the model is trained in the context of a single language and tested on the same language using 5-fold cross-validation.

4 Experimental Results

In this section, we highlight the results of our experiments regarding model robustness on out-of-domain data (Cross-Domain analyses), across multimodal inputs using various text embeddings (Cross-Modality analyses), and multilingual inputs (Cross-Language analyses).

[3] `http://cocodataset.org/#home`

"RT fluoridated water now linked to diabetes & lowered IQ still drinking it"

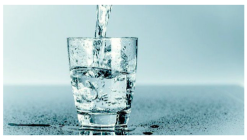

Figure 1: Incorrect classification with high confidence example: a tweet annotated *Conspiracy* classified as *Clickbait* with 92.6% confidence by the M_4 (ELMo embeddings) model.

		Predicted Class			
		Trustworthy	Deceptive	Trustworthy	Deceptive
Train on ***Twitter***	Trustworthy	46% (1.72k)	54% (2.02k)	89% (4.77k)	11% (573)
	Deceptive	19% (1.38k)	81% (5.79k)	90% (4.72k)	10% (517)
Train on ***Reddit***	**True Class** Trustworthy	87% (3.27k)	13% (469)	62% (3.28k)	38% (2.06k)
	Deceptive	89% (6.41k)	11% (757)	51% (2.69k)	49% (2.55k)
Train on ***Twitter + Reddit***	Trustworthy	60% (2.26k)	40% (1.49k)	59% (3.22k)	41% (2.21k)
	Deceptive	20% (1.43k)	80% (5.75k)	28% (1.55k)	72% (3.95k)
		Test on ***Twitter***		*Test on* ***Reddit***	

Figure 2: Confusion matrices for cross-domain analyses using models that classify content as Trustworthy or Deceptive. Cells are shaded by the rate of (mis)classification; Correct predictions are shaded in blue, misclassifications in red.

Before we do so, we manually validate the models qualitatively by examining incorrectly classified tweets with high confidences. For example, the 4-way classification model, M_4, that relies on pretrained ELMo embeddings shows the best performance, but makes errors that resemble human mistakes. The tweet shown in Figure 1 is an example of a high-confidence misclassification that could intuitively be similarly made by a human annotator. From the text or image alone the true classification is not immediately clear or intuitive from a human standpoint. The text implies conspiracy, but it is not obvious from the image – a stock image of a glass of water being poured – to which class it should belong. This analysis of high-confidence incorrect classifications provides a better understanding of how high confidence misclassifications (especially high-confidence which can serve as a proxy for where end-users *should* – or would feel that they should – trust the model) occur and under what conditions.

4.1 Cross-Domain Performance

We present the confusion matrices for the cross-domain experiments in Figure 2 and, as expected, we observe the best individual performance when models are trained and tested on same-domain (same-platform) data. Additionally, we see that performance suffers on models tested on out-of-domain data (models trained on Twitter but test on Reddit and vice-versa). However, we find that both of the cross-domain models (M_1) trained on either the Twitter or Reddit data alone over-predict the *trustworthy* class when tested on out-of-domain test data. This is irrespective of the distribution of class instances (shown in Table 2) – whether it is imbalanced (as it is in Twitter) or not (as in Reddit). As illustrated in the final row of confusion matrices, we see that the best overall performance is achieved by the model with *domain diverse* training data – trained on both the Twitter and Reddit training datasets combined. This model *performs comparably on both the Twitter and Reddit held out test sets and as well or better than the individually trained models when tested on in-domain samples.*

Our findings confirm that data from new domains (*i.e.,* social platforms) has a high impact on model performance even in the cases where classes are defined in the same way (as is the case with our Twitter and Reddit datasets that rely on annotations of the same pool of news sources as trustworthy or deceptive).

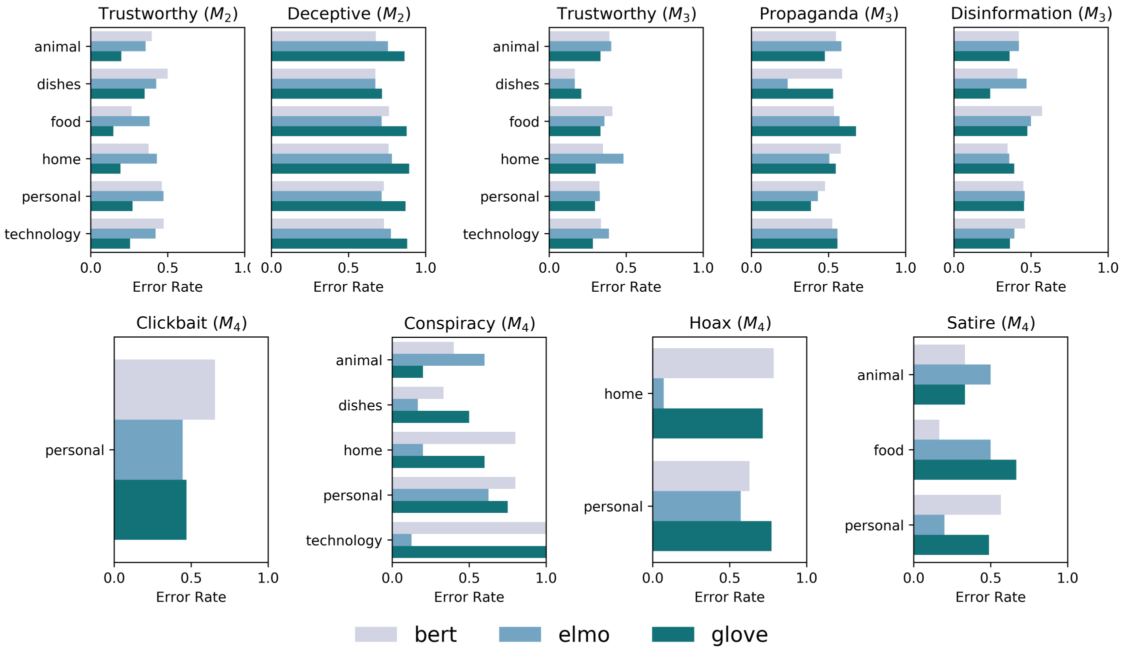

Figure 3: Error rates for Cross-Modality models for each of the six categories of primary COCO objects detected where there were at least five true class examples for which to calculate error rates.

On the Twitter test data, the domain-diverse model is even able to more accurately identify trustworthy content (60% true positives compared to 46% when trained on Twitter data alone) and reduce the number of false positive predictions of Deceptive content (40% of trustworthy tweets classified as deceptive compared to 54%). This indicates that even if used only for a single platform, engaging diverse training data examples across multiple platforms can result in a more robust, more trustworthy model.

4.2 Cross-Modality Performance

Next, we examine the performance of our multi-modal models, presented in Figure 3, across each of the three tasks and two modalities: different state-of-the-art text embeddings used by the model to represent the text component of the news post and the most distinguishable objects present in the image component of the news post to be classified. We observe several significant findings related to these modality-specific features. Because of the numerous COCO (Lin et al., 2014) classes, we group similar objects into super-categories[4]. For example, the *personal* category contains people, umbrellas, backpacks, etc.

When we examine the performance of the binary cross-modality (M_2) model, we see the best model performance on trustworthy tweets – with about 50% fewer misclassifications than on deceptive tweets – consistently across every object class and embedding type. In particular, the model that utilizes the pre-trained GloVe embeddings achieves the lowest error rate of 19.35%. We also find that the model using the GloVe embeddings has the best performance — the lowest error rate — among the 3-way (M_3) models with an error rate of 23.53%. Figure 4 illustrates an example tweet where the GloVe-based (M_3) model has correctly classified the tweet as disinformation but the models that rely on the BERT and ELMo text embeddings to represent the text component of the tweet incorrectly classify it as a propaganda tweet.

We find that Disinformation tweets where images contain food items have a misclassification rate greater than 50% for 3-way classification models (M_3) – the highest of all objects. In contrast, trustworthy tweets have the lowest misclassification error of any class across the model types for all objects at 32.15%. However, one anomaly is the rate of misclassification (48.19%) of trustworthy tweets with home-related objects (such as books) from the ELMo (M_3) model. An example of such a tweet where this ELMo model is incorrect while the others are correct is shown in Figure 5. In all three classes, we see the lowest misclassification error when kitchen-related objects are present in the image. In particular, propaganda tweets, where kitchen and personal objects appear in the image, are correctly classified

[4] https://tech.amikelive.com/node-718/what-object-categories-labels-are-in-coco-dataset/

“death by hawthorn the bath lotion that has killed over 70 russians”

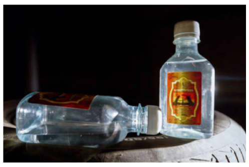

Figure 4: An example of a disinformation tweet that the GloVe (M_3) model is able to correctly identify (55.17% confidence) but the BERT and ELMo (M_3) models misclassify as propaganda (confidences of 48.29% and 60.86%).

“rt a reading society embraces civilized values is adaptable”

10-year National Policy for Reading

- UAE's multi-pronged initiatives to make reading a part of the UAE's core values and national identity.

INTEGRATING READING AT EVERY LEVEL

UAE law on reading will be announced before end of year

A Dh100million fund to support reading and its related activities will also be kicked off.

Every year will contain a month of reading. This year, it is October. For 2017, it is March.

An Educational Bag for new parents will be given to inform them on child rearing while providing their children with appropriate reading material.

Volunteering programmes to read for the elderly, the sick, children and the disabled will also be set up.

Public libraries will have branches at malls.

Reading material availability will be mandatory at cafes.

Number of UAE publications aims to go up from 400 to 4,000.

School assessment will focus on reading objectives. Training for teachers and educators on how to motivate students to read.

A survey [on reading habits] conducted with 12,000 UAE residents (teachers, students and others) concluded:

73% of parents do not read routinely.

The number of visits to public libraries represents 7% of the population whereas in Canada it is 500%.

50% of students in schools and universities do not read routinely.

Individuals read only 1.5 books a year.

78% of adults do not read regularly.

There are 20 books on average per household in comparison to 203 books per home in the UK.

Only 4 books are read every year by students in comparison to 40 in South Korea.

©Gulf News

Figure 5: The text and image from a trustworthy tweet. The ELMo model misclassifies (as disinformation with 85.74% confidence) while the BERT and GloVe models correctly classify it as trustworthy (with confidences of 76.55% and 44.27%).

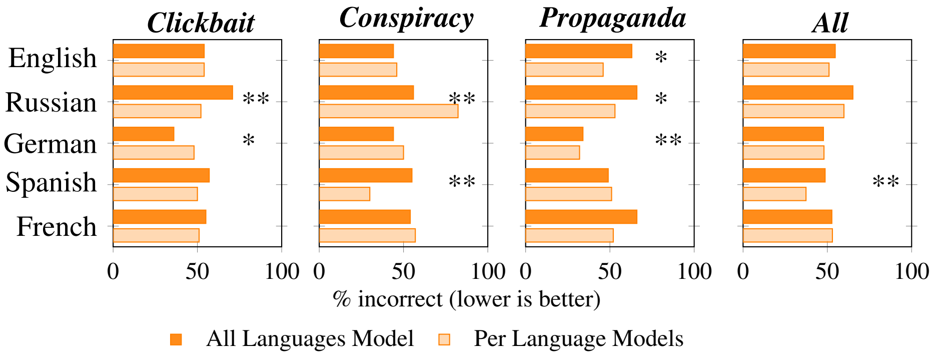

Figure 6: Bar plots display the number of incorrect predictions for each class (Clickbait, Conspiracy, Propaganda) and overall (All) for the multilingual (All Languages) model and single-language (Per Language) models over the five languages: English, Russian, German, Spanish, and French. Statistically significant differences in performance are indicated with * for $p < 0.05$ and ** for $p < 0.01$ (MWU).

more often than misclassified (an error rate less than 45%). Interestingly, the misclassification rate is similar regardless of the embedding type (between 16.67% and 20.08%) for trustworthy tweets with kitchen objects. For propaganda tweets with kitchen objects, ELMo embeddings produce the lowest misclassification rate (23.52%).

For the 4-way classification (M_4) task, we find that conspiracy tweets containing images with kitchen objects have the lowest error rate at 33.33% while conspiracy tweets with technology have the highest error rate at 70.83% compared to the remaining three deceptive classes. Predictions with ELMo embeddings have the lowest misclassification rate (36.13%) across all four classes and all object types compared to BERT (67.58%) and GloVe (61.09%). We see this particularly in cases where the BERT and GloVe models have high error rates, *e.g.,* conspiracy tweets that contain technology-related objects and hoax tweets that include home objects.

4.3 Cross-Language Performance

Finally, we illustrate the performance of the six cross-languages models in Figure 6 (in these plots, lower bars indicated higher performance): a single multilingual model trained jointly on multiple languages

(*All Languages Model*) and five models trained on each of the languages individually (*Per Language Models*, one for each of the five languages). As shown in Figure 6, the multilingual model trained jointly on multiple languages produces incorrect classifications most often on tweets that were labeled as propaganda when tested on English data while both the multilingual and single-language models tested on English data experience a steady number of errors over clickbait and conspiracy.

When we compare the jointly trained ‘All Languages’ model versus individually trained ‘Per-Language’ models across the other languages, we see that the single-language (‘Per Language’) models trained separately with English, Russian, German, and Spanish outperform the aggregated-language model tested on the same languages with regards to propaganda. The single-language model trained and tested on Spanish data displays the best performance on conspiracy. Analytically, we can see the difference in model performance when looking at each class individually and over all classes collectively. The aggregated-language model tested with German data shows better performance over clickbait and conspiracy while making more errors with propaganda. However, the single-language model trained and tested on German data sees lesser performance over clickbait and conspiracy while making fewer errors with propaganda. When we review the aggregated- and single-language German models over all classes, they display a similar performance.

5 Discussion and Future Work

We have presented an extensive evaluation of the robustness of digital deception models across frequently encountered real-world scenarios that would be necessary to benchmark against to identify reasonable performance for models considered for deployment. Our analyses have identified several trends in behavior across multiple robustness tasks and granularities of deception detection tasks (from binary to 4-way classification). To the best of our knowledge, we are the first to present this type of evaluation of deception detection model robustness.

In our robustness analyses detailed above, we have illustrated the danger of relying on single performance measurement, metrics, or analyses by showcasing how a model achieving optimal performance or significant performance increases on a specific task, domain, or context (*e.g.,* multilingual, multimodal) may significantly under-perform with slight alterations in scope or context and other frequently encountered factors in real-world applications. Further, we have shown several ways that out-of-domain, multilingual, and multiple modality inputs affect model performance and potential methods to mitigate the impact. For example, the out-of-domain analyses showed that a domain-diverse set of training examples led to higher performance on both domains in held out test sets compared to models trained on domain-specific examples. Additionally, the cross-modality analyses illustrated how frequently human-like classification mistakes can be made and which areas of the data distribution are not well represented, *e.g.,* clickbait tweets with images containing non-personal related objects. Results highlighted in the current work open several avenues of future work to develop, evaluate, and understand neural deception detection models in the context of real-world applications.

Future work will investigate the performance of a variety of such domain diverse training data compared to domain-specific models to identify if this hypothesis holds or if there may be other confounding factors related to specific platforms. In the current work, we have focused on quantifying the performance using a simplified, shared architecture that has been frequently used in deception detection models. However, future work can leverage our interactive Jupyter notebooks that will be made available at publication for reproducible extensions using our analysis framework to benchmark the performance of more complex, or newly developed state-of-the-art neural architectures.

Acknowledgements

This research was supported by the Laboratory Directed Research and Development Program at Pacific Northwest National Laboratory, a multiprogram national laboratory operated by Battelle for the U.S. Department of Energy.

References

Muhammad Aurangzeb Ahmad, Carly Eckert, and Ankur Teredesai. 2018. Interpretable machine learning in healthcare. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 559–560. ACM.

Oluwaseun Ajao, Deepayan Bhowmik, and Shahrzad Zargari. Fake news identification on twitter with hybrid cnn and rnn models. In *Proceedings of the 9th International Conference on Social Media and Society*.

Pasquale Capuozzo, Ivano Lauriola, Carlo Strapparava, Fabio Aiolli, and Giuseppe Sartori. 2020. Decop: A multilingual and multi-domain corpus for detecting deception in typed text. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1423–1430.

Weiling Chen, Yan Zhang, Chai Kiat Yeo, Chiew Tong Lau, and Bu Sung Lee. 2018. Unsupervised rumor detection based on users' behaviors using neural networks. *Pattern Recognition Letters*, 105:226–233.

Joan Claybrook and Shaun Kildare. 2018. Autonomous vehicles: No driver… no regulation? *Science*, 361(6397):36–37.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah Smith. 2019a. Show your work: Improved reporting of experimental results. *arXiv preprint arXiv:1909.03004*.

Jonathan Dodge, Q Vera Liao, Yunfeng Zhang, Rachel KE Bellamy, and Casey Dugan. 2019b. Explaining models: an empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 275–285. ACM.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE.

Maria Glenski and Tim Weninger. 2018. How humans versus bots react to deceptive and trusted news sources: A case study of active users. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE/ACM.

Maria Glenski, Tim Weninger, and Svitlana Volkova. 2018. Propagation from deceptive news sources who shares, how much, how evenly, and how quickly? *IEEE Transactions on Computational Social Systems*, 5(4):1071–1082.

Han Guo, Juan Cao, Yazi Zhang, Junbo Guo, and Jintao Li. Rumor detection with hierarchical social attention network. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*.

Christopher Glenn Harris. 2012. Detecting deceptive opinion spam using human computation. In *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Matthias Hein and Maksym Andriushchenko. 2017. Formal guarantees on the robustness of a classifier against adversarial manipulation. In *Advances in Neural Information Processing Systems*, pages 2266–2276.

Fred Matthew Hohman, Minsuk Kahng, Robert Pienta, and Duen Horng Chau. 2018. Visual analytics in deep learning: An interrogative survey for the next frontiers. *IEEE transactions on visualization and computer graphics*.

Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M Drucker. 2019. Gamut: A design probe to understand how data scientists understand machine learning models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, page 579. ACM.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. 2019. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems*, pages 9734–9745.

Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, et al. 2017. Speed/accuracy trade-offs for modern convolutional object detectors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Georgi Karadzhov, Pepa Gencheva, Preslav Nakov, and Ivan Koychev. 2017. We built a fake news & click-bait filter: What happened next will blow your mind! In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*.

Alireza Karduni, Ryan Wesslen, Sashank Santhanam, Isaac Cho, Svitlana Volkova, Dustin Arendt, Samira Shaikh, and Wenwen Dou. 2018. Can you verifi this? studying uncertainty and decision-making about misinformation using visual analytics. In *Twelfth international AAAI conference on web and social media*.

Alireza Karduni, Isaac Cho, Ryan Wesslen, Sashank Santhanam, Svitlana Volkova, Dustin L Arendt, Samira Shaikh, and Wenwen Dou. 2019. Vulnerable to misinformation?: Verifi! In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 312–323. ACM.

Srijan Kumar, Justin Cheng, Jure Leskovec, and VS Subrahmanian. 2017. An army of me: Sockpuppets in online discussion communities. In *Proceedings of the 26th International Conference on World Wide Web*, pages 857–866.

Srijan Kumar, Bryan Hooi, Disha Makhija, Mohit Kumar, Christos Faloutsos, and VS Subrahmanian. 2018. Rev2: Fraudulent user prediction in rating platforms. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 333–341. ACM.

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

Sejeong Kwon, Meeyoung Cha, and Kyomin Jung. 2017. Rumor detection over varying time windows. *PloS One*, 12(1):e0168344.

Owen Lahav, Nicholas Mastronarde, and Mihaela van der Schaar. 2018. What is interpretable? using machine learning to design interpretable decision-support systems. *arXiv preprint arXiv:1811.10799*.

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.

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.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Zachary C Lipton. 2018. The mythos of model interpretability. *Queue*, 16(3):31–57.

Xuanqing Liu, Minhao Cheng, Huan Zhang, and Cho-Jui Hsieh. 2018. Towards robust neural networks via random self-ensemble. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 369–385.

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 *Ijcai*, pages 3818–3824.

Tanushree Mitra, Graham P Wright, and Eric Gilbert. 2017. A parsimonious language model of social media credibility across disparate events. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW)*, pages 126–145. ACM.

Sina Mohseni, Eric Ragan, and Xia Hu. 2019. Open issues in combating fake news: Interpretability as an opportunity. *arXiv preprint arXiv:1904.03016*.

W James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. 2019. Interpretable machine learning: definitions, methods, and applications. *arXiv preprint arXiv:1901.04592*.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1*, pages 309–319. 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.

Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. 2018. Manipulating and measuring model interpretability. *arXiv preprint arXiv:1802.07810.*

Vahed Qazvinian, Emily Rosengren, Dragomir R Radev, and Qiaozhu Mei. 2011. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1589–1599.

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 2921–2927.

Bhavtosh Rath, Wei Gao, Jing Ma, and Jaideep Srivastava. 2017. From retweet to believability: Utilizing trust to identify rumor spreaders on twitter. *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM).*

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.

Julio Reis, André Correia, Fabrício Murai, Adriano Veloso, and Fabrício Benevenuto. 2019. Explainable machine learning for fake news detection. In *Proceedings of the 10th ACM Conference on Web Science*, pages 17–26. ACM.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144. ACM.

Greg Ridgeway, David Madigan, Thomas Richardson, and John O'Kane. 1998. Interpretable boosted naïve bayes classification. In *KDD*, pages 101–104.

Victoria L Rubin, Niall J Conroy, Yimin Chen, and Sarah Cornwell. 2016. Fake news or truth? Using satirical cues to detect potentially misleading news. In *Proceedings of NAACL-HLT*, pages 7–17.

Cynthia Rudin. 2018. Please stop explaining black box models for high stakes decisions. *arXiv preprint arXiv:1811.10154.*

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36.

Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019a. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.*

Kai Shu, Suhang Wang, and Huan Liu. 2019b. 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*, pages 312–320. ACM.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, and Huan Liu. 2020. Hierarchical propagation networks for fake news detection: Investigation and exploitation. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 626–637.

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, pages 647–653.

Svitlana Volkova, Ellyn Ayton, Dustin L Arendt, Zhuanyi Huang, and Brian Hutchinson. 2019. Explaining multimodal deceptive news prediction models. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 659–662.

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

Eric Wallace, Jens Tuyls, Junlin Wang, Sanjay Subramanian, Matt Gardner, and Sameer Singh. 2019. Allennlp interpret: A framework for explaining predictions of nlp models. *arXiv preprint arXiv:1909.09251.*

William Yang Wang. 2017. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648.*

Fan Yang, Shiva K Pentyala, Sina Mohseni, Mengnan Du, Hao Yuan, Rhema Linder, Eric D Ragan, Shuiwang Ji, and Xia Ben Hu. Xfake: Explainable fake news detector with visualizations. In *The World Wide Web Conference*.

Qiang Zhang, Aldo Lipani, Shangsong Liang, and Emine Yilmaz. 2019. Reply-aided detection of misinformation via bayesian deep learning. In *WWW*, pages 2333–2343. ACM.

Stephan Zheng, Yang Song, Thomas Leung, and Ian Goodfellow. 2016. Improving the robustness of deep neural networks via stability training. In *Proceedings of the ieee conference on computer vision and pattern recognition*.

Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. Safe: Similarity-aware multi-modal fake news detection. *arXiv preprint arXiv:2003.04981*.

Elena Zotova, Rodrigo Agerri, Manuel Núñez, and German Rigau. 2020. Multilingual stance detection in tweets: The catalonia independence corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1368–1375.

Arkaitz Zubiaga, Elena Kochkina, Maria Liakata, Rob Procter, Michal Lukasik, Kalina Bontcheva, Trevor Cohn, and Isabelle Augenstein. 2018. Discourse-aware rumour stance classification in social media using sequential classifiers. *Information Processing & Management*, 54(2):273–290.