Evaluating Deception Detection Model Robustness To Linguistic Variation

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

Abstract

With the increasing use of machine-learning driven algorithmic judgements, it is critical to develop models that are robust to evolving or manipulated inputs. We propose an extensive analysis of model robustness against linguistic variation in the setting of *deceptive news detection*, an important task in the context of misinformation spread online. We consider two prediction tasks and compare three state-of-the-art embeddings to highlight consistent trends in model performance, high confidence misclassifications, and high impact failures. By measuring the effectiveness of adversarial defense strategies and evaluating model susceptibility to adversarial attacks using character- and word-perturbed text, we find that character or mixed ensemble models are the most effective defenses and that character perturbation-based attack tactics are more successful.

1 Introduction

Over two-thirds of US adults get their news from social media, but over half (57%) "expect the news they see on social media to be largely inaccurate" (Shearer and Matsa, 2018). A 2020 Reuters Institute global news survey found a similar trend with 56% of respondents concerned with misinformation in online news (Newman et al., 2020). There are online and offline impacts from the spread of misinformation or deceptive news stories within online communities. However, the rate at which new content is submitted to social media platforms is a significant obstacle for approaches that require manual identification, annotation, or intervention. In recent efforts, evaluation has focused on aggregate performance metrics on test sets often collected from social media platforms like Twitter, Facebook, or Reddit (Rubin et al., 2016; Mitra et al., 2017; Wang, 2017) but these platforms are not representative in regards to user demographics or topics of discussion. Further, aggregate performance metrics are not sufficient to provide insight on generalizable performance.

When we consider the identification of deceptive news online — where humans often disagree on or challenge the judgements of others (Karduni et al., 2018, 2019; Ott et al., 2011) — we need more rigorous evaluations of model decisions, with a focus on expected performance across varied or manipulated inputs. Our work examining reliability of performance when faced with linguistic variations is a step towards comprehensively understanding model robustness that may highlight inequalities in cases of failure. Although machine learning models are often leveraged for their ability to tackle rapid response at scale, it is critical to understand nuanced model biases and the significant downstream consequences of model decisions on users.

A known gap exists in our understanding of underlying machine learning decision-making processes, particularly with deep learning "black-box" models. The use of traditional, aggregate metrics for model performance, such as accuracy or F1 score, are not sufficient in pursuit of this understanding. We argue that evaluations need to explicitly measure the extent to which model performance is affected by data with a varied topic distribution. Evaluations highlighting when models are correct, which examples can provide explanations, and clarification or reasoning for why a user should trust a given model are well-aligned with recent themes in research on machine learning interpretability, trust, fairness, accountability, and reliability (Lipton, 2018; Doshi-Velez and Kim, 2017; Hohman et al., 2018).

In this paper, we perform an adversarial model evaluation across two multimodal deception prediction tasks to identify which defensive strategies are most successful across a variety of attacks. Our main contribution is a framework of analysis for model robustness across variations in linguistic sig-

Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media, pages 70–80
Online Workshop, June 10, 2021. ©2021 Association for Computational Linguistics

nals and representations that may be encountered in real-world applications of digital deception models (*e.g.,* natural linguistic differences, evolving tactics from deceptive adversaries to evade detection). In particular, we present evaluations on the susceptibility of widely used text embeddings to naive adversarial attacks, which types of text perturbations lead to the most high-confident errors, and to what extent our findings are task specific. The perturbed text emulates real examples of linguistic variations, e.g. non-native speakers, spelling mistakes, or shortened online speech. Our evaluations reveal how models react to perturbed text which we argue is a likely occurrence when deployed in a real-world setting.

2 Related Work

With the increasing concern for the impact of misinformation and deceptive news content online, many studies have explored or developed models that detect such news. Recent efforts focus on identifying a spectrum of deception: from binary classification of content as suspicious and trustworthy (Volkova et al., 2017) to a more fine-grained separation within deceptive classes (*e.g.,* propaganda, hoax, satire) (Rashkin et al., 2017). Additional work has explored the behavior of malicious users and bots (Glenski and Weninger, 2018; Kumar et al., 2017, 2018) and spread patterns of misinformation or rumors (Kwon et al., 2017; Vosoughi et al., 2018) to aid in classification tasks. Strong evidence suggests that enriched features such as images, temporal and structural attributes, and linguistic features boost model performance over dependence on textual characteristics alone (Wang, 2017; Qazvinian et al., 2011; Kwon et al., 2013). The need for effective, trustworthy, and interpretable detection models is a vital concern and must be an essential requirement for models where decisions or recommendations can significantly affect end users.

A variety of deep learning architectures applied to deception detection tasks include convolutional neural networks (CNNs) (Ajao et al., 2018; Wang, 2017; Volkova et al., 2017), long short-term memory (LSTM) models (Chen et al., 2018; Rath et al., 2017; Zubiaga et al., 2018; Zhang et al., 2019) , and LSTM variants with attention mechanisms (Guo et al., 2018; Li et al., 2019). Architecture and other aspects of neural network design typically depend on the classification task and require specialized hyperparameter tuning. In order to provide a fair comparison of model evaluations across tasks and for the purpose of consistency across experiments, we implement a multimodal LSTM model similar to recent work. Our approach allows for more accurate comparisons of factors related to adversarial susceptibility across classification tasks. Developing novel state-of-the-art models for deception detection or comparing multiple architectures is beyond the scope of this paper.

Although popularly used across many domains, deep learning systems can be extremely brittle when evaluated on examples outside of the training data distribution (Goodfellow et al., 2014; Moosavi-Dezfooli et al., 2017; Fawzi et al., 2018). Nguyen *et al* (2015) have shown that small perturbations in input data can cause highly probable misclassifications. Further research demonstrates additional attacks that make neural networks more susceptible to adversaries such as locally trained DNNs to crafted adversarial inputs (Papernot et al., 2016c,a) and gradient-based attacks (Biggio et al., 2013). To counteract these offensive strategies, proposed methods of defense include augmented training data with adversarial examples (Tramèr et al., 2018), training a separate model to distinguish genuine data from malicious data (Metzen et al., 2017), and implementing a defensive distillation mechanism to increase a model's resiliency to data poisoning (Papernot et al., 2016b). However, as defense strategies are created, new attacks are continually developed to circumvent them (Carlini and Wagner, 2017). While there is a focus on image perturbations and related attacks, textual data is similarly vulnerable to such strategies (Gao et al., 2018; Samanta and Mehta, 2017; Liang et al., 2018)). The susceptibility of deception detection models to text-based adversarial attacks as well as the effectiveness of defense strategies have not been extensively evaluated.

3 Methodology

In this section, we introduce our detection tasks, models, and evaluation methods. We randomly perturb words or characters with their nearest neighbors to mimic a low-effort adversarial attack (*e.g.,* replacing words with synonyms) as opposed to methods that assume an adversary has technical expertise or require sophisticated augmentations (*e.g.,* gradient-based algorithms). We argue that robustness against these low-effort attacks is a necessary first step towards trustworthy models; these attacks are reflective of natural or unintentional variations (*e.g.,* misspellings, non-native speaker discussions) as well as sophisticated strategies.

3.1 Deception Detection Tasks

We apply a comprehensive evaluation of model robustness and susceptibility to two classification tasks[1]: 3-way (trustworthy, propaganda, disinformation) and 4-way (clickbait, hoax, satire, conspiracy). Including both allows us to compare defense and attack strategies across models at varied levels of deception and evaluate method generalizability.

The 3-way task includes two extreme deceptive classes, propaganda and disinformation, and seeks to differentiate them from "trustworthy" sources (Derakhshan and Wardle, 2017). Due to the stronger intent to deceive of these classes, we expect a model to distinguish trustworthy news more easily and expect more confusion when classifying news as either propaganda or disinformation. Misclassifications of these as trustworthy will have a greater negative impact. To better identify high-impact errors, we collapse disinformation and propaganda into a single class as part of a binary sub-task separating trustworthy from deceptive.

The 4-way task centers lesser deceptive content, the classes included have a lower intent to deceive and are more difficult to distinguish from one another. For instance, satirical news sites produce humorous content or social commentary rather than deliberately false information and have a low intent to deceive audiences (Fletcher and Nielsen, 2017). Because of this inherent difference from the other deceptive news types, we include a binary sub-task separating satire from the remaining classes.

3.2 Data Collection and Annotation

Models were trained and tested on Twitter API data. Our corpus comprises English retweets with images from official news media Twitter accounts. Class labels are based on "verified" news sources and a public list of sources annotated along the spectrum of deceptive content (Volkova et al., 2017)[2] from 2016. Thus, we limit our corpus to that 12 month period of activity. The 3-way and 4-way task data consist of 54.5k and 2.5k tweets. Although there are limits to source-level annotations (*e.g.,* tweets of different deceptive classes shared from a single source), we advocate for focus on news sources rather than individual stories, similar to previous work (Vosoughi et al., 2018; Lazer et al., 2018). We posit the definitive element of deception to be the intent and tactics of the source.

3.3 Multimodal Deception Detection Models

We clean the tweet text by lowercasing and removing punctuation, mentions, hashtags, and URLs. We encode biased and subjective language as frequency vectors constructed from LIWC (Pennebaker et al., 2001) and several lexical dictionaries such as hedges and factives (Recasens et al., 2013) which are often used for text classification (Rashkin et al., 2017; Shu et al., 2019).

We implement a two-branch architecture[3] that leverages text, lexical features, and images. The text branch consists of a pre-trained text embedding layer, an LSTM layer, and a fully connected layer. The output is concatenated to the lexical feature vector before being passed to another fully connected layer. In the second branch, we pass the image vector through a fully connected, two layer network. The combined text embeddings and lexical features are concatenated with the processed image representation which is then fed to a fully connected network for classification. Our chosen architecture resembles current systems in deployment and allows us to complete complex analyses.

3.4 Model Evaluation Methods

We perform a comprehensive evaluation for both tasks over *embeddings*, *defenses*, and *attacks*. This section describes our text perturbation methods and our defense and attack frameworks.

3.4.1 Varying Text Representations

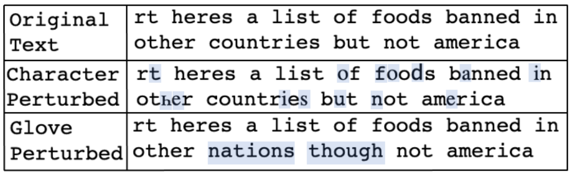
We consider three embedding techniques that have shown state-of-the-art performance on several NLP tasks: GLoVe (Pennington et al., 2014), ELMo (Peters et al., 2018), and BERT (Devlin et al., 2019). We recognize that each embedding method was trained on separate data[4], under different conditions, and produces various sized vectors. Thus, we fine-tune the embedding layer during training.

[1] Although we chose to use these two tasks, our framework is task-agnostic and can be applied to any classification task.

[2] www.cs.jhu.edu/~svitlana/data/SuspiciousNewsAccountList.tsv; www.cs.jhu.edu/~svitlana/data/VerifiedNewsAccountList.tsv

[3] Parameters selected by a random search: Adam optimizer, 10^{-6} learning rate, 0.2 drop out, and 10 training epochs.

[4] We use GloVe (Twitter 27B), ELMo (tfhub.dev/google/elmo/2), and BERT (github.com/huggingface/transformers)



Original Text	rt heres a list of foods banned in other countries but not america
Character Perturbed	rt heres a list of foods banned in otнer countries but not america
Glove Perturbed	rt heres a list of foods banned in other nations though not america

Figure 1: Examples of adversarial perturbations.

3.4.2 Linguistic Variation

We examine how changes to text input affect model performance using character and word perturbations and focus on the impact of naive linguistic variations in text. For character-level perturbations, we randomly replace 25% of characters in each tweet with a Unicode character that is indistinguishable from the original to a human (as shown in Figure 1). This approach, known in computer security as a homograph attack or script spoofing, has been investigated to identify phishing or spam (Fu et al., 2006b,a; Liu and Stamm, 2007) but has not been applied in the NLP domain to our knowledge. For word-level perturbations we randomly replace 25% of words with a nearest neighbor in the each embedding space using Annoy[5].

3.4.3 Defense Viewpoint

To evaluate the efficacy of common defenses to guard against adversarial attacks – augmenting the training data – we perturb our training set (Tr) to varying degrees using each linguistic variation strategy. We compare the following defenses:

- Tr: train with original examples;
- $Tr^{50\%}$: train with half of the examples perturbed;
- $Tr^{'}$: train with all examples perturbed;
- Ensemble (E): majority vote of ensemble of models trained on Tr, $Tr^{50\%}$, and $Tr^{'}$.

Each defense has been perturbed for each embedding type. For example, we train our models using four variations of $Tr^{50\%}$: $Tr^{50\%}_C$ (50% of examples perturbed using the character-level attack), and three $Tr^{50\%}_W$ defenses with 50% of the examples perturbed using the word-level attack ($Tr^{50\%}_{BERT}$, $Tr^{50\%}_{ELMo}$, $Tr^{50\%}_{GLoVe}$).

We use three sets of ensembles: (1) E_C, an ensemble of models trained with Tr, $Tr^{50\%}_C$, and $Tr^{'}_C$, (2) E_W, an ensemble of models trained with Tr, $Tr^{50\%}_W$, and $Tr^{'}_W$, and (3) E_{C+W}, an ensemble of five models trained on Tr, $Tr^{50\%}_C$, $Tr^{'}_C$, $Tr^{50\%}_W$, and $Tr^{'}_W$. Higher confidence predictions are used in the case of ties.

We test the performance of models trained using each defense on fully perturbed ($Te^{'}$) and the original, unperturbed test data (Te). Ideally, we want models to perform well on both so we also consider three $Mixed$ test sets (Te + $Te^{\prime}_C$ + $Te^{\prime}_W$), one for each $Te^{'}_W$ ($Mixed_{BERT}$, $Mixed_{ELMo}$, and $Mixed_{GLoVe}$).

3.4.4 Attack Viewpoint

We also evaluate the impact of the linguistic perturbations as adversarial attack strategies. The attack test sets were perturbed similarly to the train sets:

- Te: original examples (no attack);
- $Te^{'}_C$: all examples perturbed (char-level);
- $Te^{'}_W$: all examples perturbed (word-level).

As with the defense viewpoint, we have four sets of the $Te^{'}$ test data used to evaluate each attack condition: $Te^{'}_C$, $Te^{'}_{BERT}$, $Te^{'}_{ELMo}$, $Te^{'}_{GLoVe}$.

3.4.5 High Confidence And High Impact

For researchers and end-users to establish trust in the models they develop or use, it is essential to understand the circumstances in which a model would make a highly confident misclassification. Inherently, model confidence measures the certainty of a prediction and quantifies the expertise and stability of a model. We closely examine instances in which our models have incorrectly predicted the class of a tweet with high confidence (greater than 90%) to identify potential weaknesses of the models.

Traditional performance metrics (F1 score, precision, recall) treat misclassifications with high confidence and low confidence alike. While overall error is an important measure, a model with a slightly higher overall failure rate but lower confidence may result in a better "worst case" outcome if appropriately incorporated in a semi-automated or human-in-the-loop deployment strategy that considers the uncertainty of predictions or recommendations via the model confidence before taking action.

We also examine high impact errors using the binary sub-tasks for each classification task as described above. In this analysis, we identify how often models make significant errors. For example, mistaking a post labeled as disinformation for trustworthy (an opposite class) rather than propaganda (a similar class).

4 Experimental Results

In this section, we detail our results when evaluating different combinations of adversarial defenses

[5] https://github.com/spotify/annoy

3-way	**Character** ($\Delta Te'_C$)			**Word** ($\Delta Te'_W$)		
Defense	*BERT*	*ELMo*	*GloVe*	*BERT*	*ELMo*	*GloVe*
Tr	+36%	+34%	+33%	+37%	+37%	+38%
$Tr^{50\%}$	+2%	-2%	-2%	-1%	-3%	-3%
$Tr^{'}$	+1%	-1%	-1%	-6%	-21%	-5%
E_C	+2%	+4%	+7%	-5%	-8%	-0%
E_W	+14%	+12%	+15%	+11%	+21%	+17%
E_{C+W}	+10%	+7%	+7%	+7%	+9%	+3%
4-way	**Character** ($\Delta Te'_C$)			**Word** ($\Delta Te'_W$)		
Defense	*BERT*	*ELMo*	*GloVe*	*BERT*	*ELMo*	*GloVe*
Tr	+14%	+52%	+0%	+5%	+53%	+6%
$Tr^{50\%}$	+32%	+31%	+36%	+16%	+10%	+16%
$Tr^{'}$	+30%	+30%	+32%	-7%	-3%	-5%
E_C	+11%	+15%	+2%	+13%	+17%	+10%
E_W	+28%	+48%	+21%	+8%	+25%	+6%
E_{C+W}	+17%	+21%	+22%	+17%	+9%	+15%

Fewer errors on Te' 🟦🟦🟦 More errors on Te' 🟧🟧🟧

Table 1: Relative difference in error rate for each task's perturbed test data ($Te^{'}$) compared to original Te.

and attacks. In order to produce a holistic evaluation of model susceptibility, we examine defenses and attacks separately. Although we consider the same model behavior, each position can highly impact the interpretation of the findings and key takeaways. We also want to understand model misclassifications, including those with high model confidence and those that can have a greater negative effect in practice which we accomplish with our high confidence and high impact analyses.

4.1 Defense Viewpoint

We compare results from the models trained on data with varying degrees of perturbation to understand which models provide the most effective defenses. We define success in the defender case as the lowest error rate across a variety of test data including original (Te), perturbed ($Te^{'}$), and combinations of original and perturbed samples ($Mixed$). We start by presenting the *relative difference* in error rates which is the percentage increase or decrease in the error rate of the perturbed ($\hat{Te}^{'}$) and original ($\hat{Te}$) test data. Relative difference is defined as:

$$\Delta Te'_x = \frac{\hat{Te}^{'}_x - \hat{Te}}{\hat{Te}} \qquad (1)$$

where x represents perturbation type (char or word). Relative difference results are shown in Table 1.

With the 3-way task, defenses across embeddings and test data appear effective and achieve low relative percent differences with the exception of models trained with the original examples (Tr). The Tr defense is ineffective against both the character- and word-perturbed text ($Te^{'}_C$ and $Te^{'}_W$). Intuitively, this could be seen as an "out of domain" data attack where the perturbed test set has significantly changed the original distribution such that a model not trained on perturbed data is more susceptible to errors. The E_C models have a lower relative difference in errors on $Te^{'}_W$ than on $Te^{'}_C$ across all three embeddings used for text representations. Thus, an ensemble of models overcomes the setback of out of domain data.

On $Te^{'}_C$ data, we observe similar relative errors between the three embeddings for all defense types; however, the performance on $Te^{'}_W$ is much more varied with the largest change seen from the ELMo embeddings, -1% relative difference from the $Tr^{'}$ model on $Te^{'}_C$ and -21% relative difference from the same defense on $Te^{'}_W$. We only see consistent behavior with the $Tr^{50\%}$ defense when tested on Te and $Te^{'}$ across embedding strategies and attack perturbations. A model trained on data containing 50% clean and 50% perturbed samples performs almost equally on the clean and perturbed test sets and exhibits less than a 5% difference in errors between the test sets for all embeddings.

Dissimilarly, the 4-way task defenses display higher relative differences in errors on $Te^{'}_C$ and $Te^{'}_W$ with the exception of the $Tr^{'}$ defense. Under the $Te^{'}_W$ attack, $Tr^{'}$ is the only defense to achieve fewer errors on the perturbed test set. We also see more variation in the relative errors across embeddings for the same defenses. For instance, with BERT, the Tr model defending against the $Te^{'}_C$ attack has a 14% relative error difference while the equivalent ELMo and GloVe models have 52% and 0% relative error differences, respectively. This trend appears across defenses and in some cases highlights the ineffectiveness of these defenses.

With both tasks, there are fewer errors on $Te^{'}_W$ using $Tr^{'}$ as the defense, regardless of embedding type, specifically 21% fewer errors on the 3-way task and 3% fewer errors on the 4-way task with ELMo embeddings. Although the results on the tasks look dissimilar in terms of “best generalizability" (*i.e.,* show good performance on both Te and $Te^{'}$), we see that character-based ensemble models exhibit the most consistent defense across tasks. The ability to have a single model (E_C) perform uniformly well across tasks outweighs the slight performance increase with individualized models per task. The ensemble defenses that leverage character-based defenses (E_C or E_{C+W})

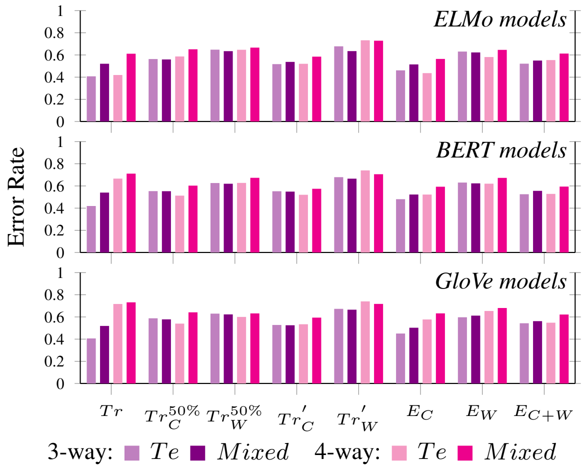


Figure 2: Defense effectiveness illustrated by error rate as a function of defense strategy for each model when tested on Te or $Mixed$ (Te + Te'_C + Te'_W) data.

are more generalizable to novel test data which is beneficial when considering real-world data.

Performance on a variety of test data alone does not indicate the best defense. If a given defense performs similarly across datasets, it may simply perform equally poorly. Pairing additional analysis shown in Figure 2 with generalizability results highlighted in Table 1, we can better investigate effective defenses. In Figure 2, we plot the error rates of each defense when paired with clean (Te) or a mixed combination of clean and poisoned (Te + Te'_C + Te'_W) examples. While the ELMo Tr models in Table 1 had the highest relative differences in error, these models outperform the same BERT and GloVe models. The best defenses (*i.e.,* with the lowest error rates) are the same models that were most consistently generalizable across attack types – E_C and E_{C+W}. *These results indicate that defenses that include character-perturbed training data (E_C and E_{C+W}) are the most effective against character- and word-based attacks.*

4.2 Attack Viewpoint

Next, we examine susceptibility to adversarial attacks from the view of the attacker. We consider the impact on model confidence, and we analyze how a given attack impacts the uncertainty of classifications overall. For example, in a human-machine teaming scenario, a deception detection model would be used to flag content for a human fact-checker who may rely on the model's confidence when choosing whether to trust the classification.

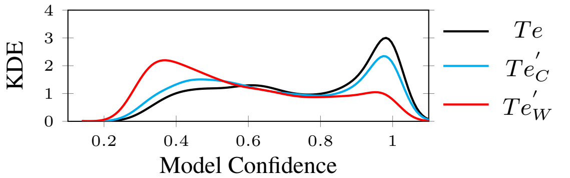


Figure 3: Kernel density estimation (KDE) plots illustrating distribution of model confidences.

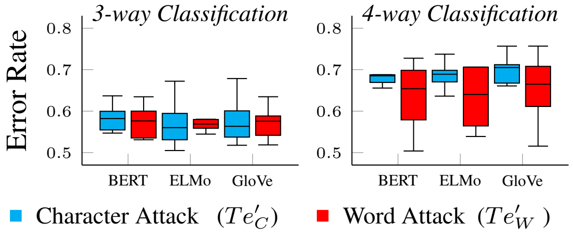


Figure 4: Box plots showing the effectiveness of the character and word perturbation attack tactics via error rates across BERT-, ELMo-, and GloVe-based models.

In Figure 3, KDE plots illustrate model confidence distributions across examples from three test sets. We find that Te'_W peaks at lower model confidences and flattens out as model confidence increases. By contrast, Te and Te'_C peak at a model confidence close to 1. This shows that there is more confusion for predictions made on word-perturbed test data. If an analyst or end user relies on model confidence when choosing to accept a prediction, a significant difference in uncertainty of model classification can affect that decision. For example, when testing on clean examples (Te), the shift to a lower overall confidence may be enough to degrade the efficacy of the recommendation, even if the model has correctly classified the example.

Having examined the impact of attacks on model confidence, we next compare the effectiveness of each attack tactic when success is defined by the number of misclassifications. In Figure 4, box plots show the number of misclassifications as error rates. Te'_C and Te'_W attacks achieve similar median error rates in the 3-way task, and the maximum error rates are greater for the character than the word attacks. Although the 4-way task shows more discrepancies across attacks, again we see the character attacks display larger rates of error. *With both tasks, we see the largest number of misclassifications typically result from character-based attacks.*

Of note, the impact on the 3-way task is consistent across embedding types and attacks (the median error rates range from 56% to 59%). We see the widest range and largest maximum error rate with ELMo- and GloVe-based models when attacked with character-perturbed text. Contrastly, the 4-way task displays similar trends across em-

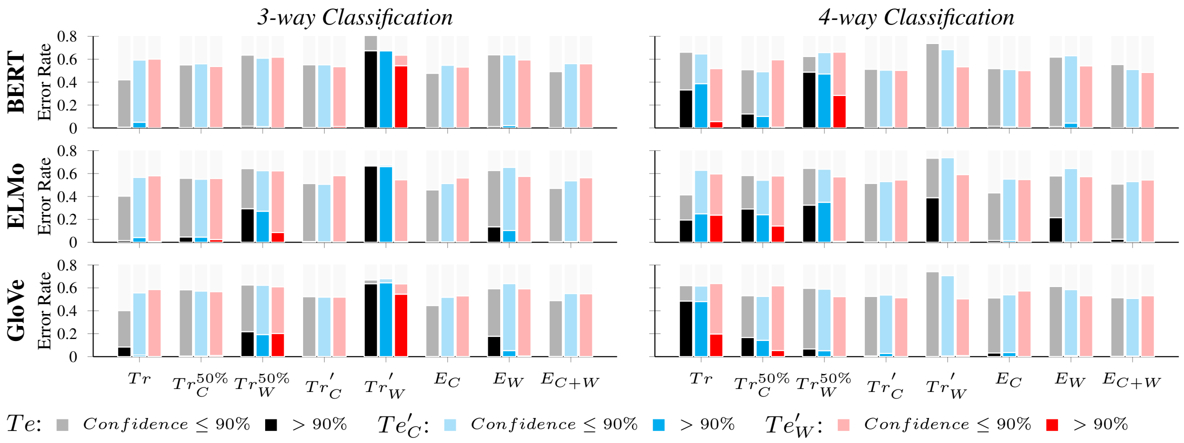


Figure 5: Error rates highlighting the prevalence of *high confidence* ($> 90\%$) errors when tested on each dataset.

beddings but not across attack types. Although we see a greater range of error rates with the word attacks, the character attacks achieve a larger median and mean error rates than either Te or $Te_W^{'}$.

4.3 High Confidence Misclassifications

Next, we examine high confidence misclassifications which are integral to understanding model behavior and the limitations faced by deceptive news detection approaches. Figure 5 highlights error rates across test data distinguishing high confidence ($> 90\%$) from lower confidence ($\leq 90\%$).

With the 3-way task, we observe that high confidence misclassifications account for a majority of all errors from the $Tr_W^{'}$ models (85.7% of errors with the $Te_W^{'}$ attack are considered high confidence). This is larger than the errors from any of the other models. We also notice one exception to this finding: the $Tr_W^{'}$ ELMo model makes very few (less than 0.5%) high confident incorrect predictions for the $Te_W^{'}$ test set. With the 4-way task, we do not see the same frequency of high confidence errors although $Tr_W^{50\%}$ displays high rates of high confidence misclassifications on BERT and ELMo models when tested with Te and $Te_C^{'}$.

Previously, we detailed stronger performance from the E_C and E_{C+W} defenses. As shown in Figure 5, both ensemble defenses display the lowest (or second lowest) error rates across attacks. Moreover, these models exhibit sparse high confidence misclassifications when reviewing averaged confidence scores across ensembled models. This is advantageous model behavior in a real-world setting when predicted model confidences must act as a proxy for uncertainty, and, in instances when ground truth labels are unknown, as a means to calibrate users' trust in model classifications.

4.4 High Impact Misclassifications

Finally, we contrast model performance for each task and our devised binary sub-tasks (trustworthy versus deceptive for the 3-way task and satire versus not satire for the 4-way task). Figure 6 demonstrates model tendencies towards high impact misclassifications across defenses, embeddings, and test sets. A higher binary F1 score indicates fewer high impact misclassifications – *i.e.,* more errors due to misclassifications among similar classes as compared to more errors due to misclassifications among significantly different classes. All models exhibit higher F1 scores on the binary sub-task than the multiclass task, as would be expected since the binary task presents an "easier" problem with an increased random chance for correct classification.

We examine consistent trends for each test set (indicated by color) or embedding type (indicated by mark size) across defenses. Values plotted in the same color cluster more consistently than those plotted in the same size. Two defenses show the most consistency in performance across configurations. $Tr_W^{50\%}$ displays low performance on both the binary and multiclass formulations of the 3-way task and $Tr_C^{'}$ displays high performance (relative to each task) across formulations for both the 3-way and 4-way tasks, with more consistency and higher performance on the 4-way task. Similar to $Tr_C^{'}$, $Tr_C^{50\%}$ displays high performance across both tasks, although this defense is more consistent on the 3-way task. Although the Tr model is the best configuration when testing on Te, the Tr model shows much lower efficacy when tested against both attacks. *Overall, configurations using*

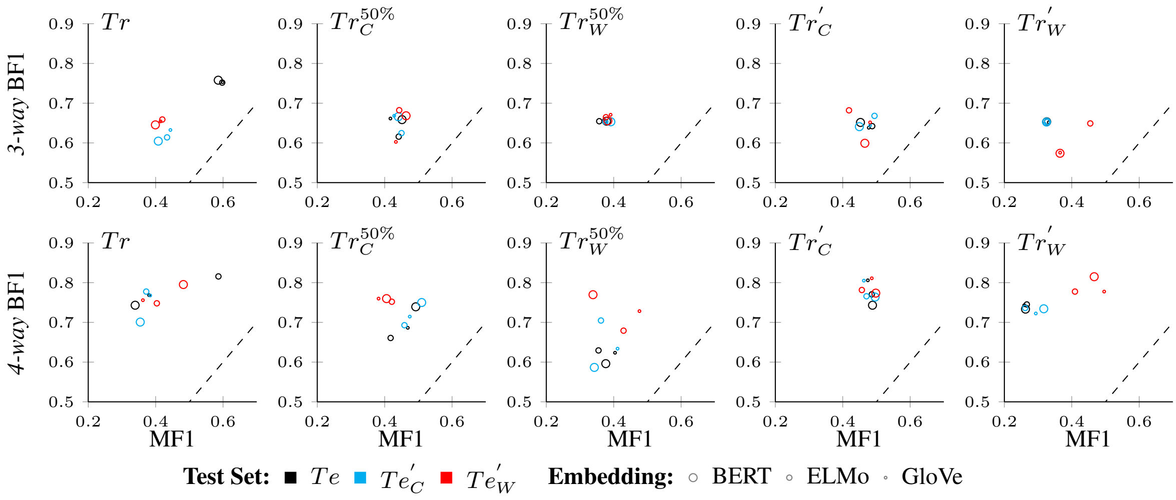


Figure 6: Binary F1 (BF1) as a function of multiclass F1 (MF1). Dashed lines indicate equal performance.

the character-based defenses result in the fewest overall high impact misclassifications.

Interestingly, we see that defenses are more effective at the binary sub-task for the 4-way classification (satire versus not satire) than the binary sub-task for the 3-way classification (trustworthy versus deceptive). Both trustworthy and deceptive news media attempt to present the information and news they share as factual, truthful content. In contrast, satire is distinct from other types of deceptive news as well as distinct from trustworthy news sources because it does not intend to present content as factual or accurate. This distinction between the classes considered in the binary sub-tasks can explain the observed difference in performance.

5 Discussion and Future Work

Linguistic variation in text (adversarial or otherwise) is frequently encountered in real-world settings. As such, we have presented extensive evaluations concerning the robustness of deception detection models to perturbed inputs. To the best of our knowledge, we are the first to evaluate model susceptibility in regards to adversarial linguistic attacks, investigate model behavior behind high confident or high impact failures, and present effective defensive strategies to these types of attacks. Our comprehensive set of perturbation experiments identify key findings from not only the defender perspective (the most effective strategy of defense across multiple or combined attacks) but also the attacker perspective (the most effective method of attack) – a focus of analysis not previously studied. In regard to the defense viewpoint, we show that ensemble-based approaches leveraging perturbed (adversarial) and non-perturbed (original) training examples perform consistently well. With the attack viewpoint, character-based attacks hinder performance regardless or model, defense, or task.

Our adversarial analyses have also illustrated the danger of relying on single performance metrics. Models that achieve optimal performance on a specific task or adversarial situation may significantly under-perform with slight alterations in scope or context. For example, although the E_C and E_{C+W} models saw second best performance on either classification task, they outperformed the “best models” when considering all possible attacks. The models with the highest overall performance were also not consistently found to have the lowest high confidence or high impact misclassifications – an important consideration if a model is being considered for use on live platforms where decisions can significantly impact users.

The results highlighted in this work provide justification for enhanced development and analysis of deception detection models. Although we rely on a consistent model architecture in order to make equitable comparisons across tasks and datasets, the evaluation framework we present can be replicated with additional models, complex architectures, and variants in test data. This work relies on uniform perturbation attacks as opposed to strategic perturbation strategies that target specific substrings – such as pseudonymous terms, phrases, or monikers. Subsequent experiments will investigate more complex strategic attacks and their ability to evade or confuse deception detection models.

Acknowledgements

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

References

Oluwaseun Ajao, Deepayan Bhowmik, and Shahrzad Zargari. 2018. Fake news identification on twitter with hybrid cnn and rnn models. In *Proceedings of the 9th International Conference on Social Media and Society*. ACM.

Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. 2013. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 387–402. Springer.

Nicholas Carlini and David Wagner. 2017. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57. IEEE.

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.

Hossein Derakhshan and Claire Wardle. 2017. Information disorder: definitions. *AA. VV., Understanding and addressing the disinformation ecosystem*, pages 5–12.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*.

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

Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. 2018. Analysis of classifiers' robustness to adversarial perturbations. *Machine Learning*, 107(3):481–508.

Richard Fletcher and Rasmus Kleis Nielsen. 2017. People dont trust news media–and this is key to the global misinformation debate. *AA. VV., Understanding and Addressing the Disinformation Ecosystem*, pages 13–17.

Anthony Y Fu, Xiaotie Deng, Liu Wenyin, and Greg Little. 2006a. The methodology and an application to fight against unicode attacks. In *Proceedings of the second symposium on Usable privacy and security*, pages 91–101. ACM.

Anthony Y Fu, Wan Zhang, Xiaotie Deng, and Liu Wenyin. 2006b. Safeguard against unicode attacks: generation and applications of uc-simlist. In *Proceedings of WWW*.

Ji Gao, Jack Lanchantin, Mary Lou Soffa, and Yanjun Qi. 2018. Black-box generation of adversarial text sequences to evade deep learning classifiers. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 50–56. 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 ASONAM*. IEEE/ACM.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Han Guo, Juan Cao, Yazi Zhang, Junbo Guo, and Jintao Li. 2018. Rumor detection with hierarchical social attention network. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. ACM.

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*.

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 IUI*. ACM.

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 *Proceedings of ICWSM*.

Srijan Kumar, Justin Cheng, Jure Leskovec, and VS Subrahmanian. 2017. An army of me: Sockpuppets in online discussion communities. In *Proceedings of WWW*, 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 ACM WSDM*. ACM.

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

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.

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.

Bin Liang, Hongcheng Li, Miaoqiang Su, Pan Bian, Xirong Li, and Wenchang Shi. 2018. Deep text classification can be fooled. In *Proceedings of IJCAI*.

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

Changwei Liu and Sid Stamm. 2007. Fighting unicode-obfuscated spam. In *Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit*, pages 45–59. ACM.

Jan Hendrik Metzen, Tim Genewein, Volker Fischer, and Bastian Bischoff. 2017. On detecting adversarial perturbations. *arXiv preprint arXiv:1702.04267*.

Tanushree Mitra, Graham P Wright, and Eric Gilbert. 2017. A parsimonious language model of social media credibility across disparate events. ACM CSCW.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. 2017. Universal adversarial perturbations. In *Proceedings of IEEE CVPR*, pages 1765–1773.

Nic Newman, Richard Fletcher, Anne Schulz, Simge Andı, and Rasmus Kleis Nielsen. 2020. Reuters institute digital news report 2020. *Reuters Institute*.

Anh Nguyen, Jason Yosinski, and Jeff Clune. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of IEEE CVPR*, pages 427–436.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of ACL*, pages 309–319. ACL.

Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. 2016a. The limitations of deep learning in adversarial settings. pages 372–387. IEEE.

Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. 2016b. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 582–597. IEEE.

Nicolas Papernot, Patrick D. McDaniel, Ian J. Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. 2016c. Practical black-box attacks against deep learning systems using adversarial examples. *CoRR*.

James Pennebaker, Martha Francis, and Roger Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of EMNLP*. ACL.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT*.

Vahed Qazvinian, Emily Rosengren, Dragomir R Radev, and Qiaozhu Mei. 2011. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of 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 EMNLP*.

Bhavtosh Rath, Wei Gao, Jing Ma, and Jaideep Srivastava. 2017. From retweet to believability: Utilizing trust to identify rumor spreaders on twitter. In *Proceedings of ASONAM*.

Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic models for analyzing and detecting biased language. In *Proceedings of ACL*, pages 1650–1659.

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*.

Suranjana Samanta and Sameep Mehta. 2017. Towards crafting text adversarial samples. *arXiv preprint arXiv:1707.02812*.

Elisa Shearer and Katerina Eva Matsa. 2018. News use across social media platforms 2018.

Kai Shu, Suhang Wang, and Huan Liu. 2019. Beyond news contents: The role of social context for fake news detection. In *Proceedings of WSDM*, pages 312–320. ACM.

Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. 2018. Ensemble adversarial training: Attacks and defenses. In *ICLR*.

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 ACL*, volume 2, pages 647–653.

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

William Yang Wang. 2017. “liar, liar pants on fire”: A new benchmark dataset for fake news detection. In *Proceedings of ACL*, pages 422–426.

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

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.