Improving Cross-domain, Cross-lingual and Multi-modal Deception Detection

Subhadarshi Panda
Hunter College
City University of New York
spanda@gc.cuny.edu

Sarah Ita Levitan
Hunter College
City University of New York
sarah.levitan@hunter.cuny.edu

Abstract

With the increase of deception and misinformation especially in social media, it has become crucial to be able to develop machine learning methods to automatically identify deceptive language. In this proposal, we identify key challenges underlying deception detection in cross-domain, cross-lingual and multi-modal settings. To improve cross-domain deception classification, we propose to use inter-domain distance to identify a suitable source domain for a given target domain. We propose to study the efficacy of multilingual classification models vs translation for cross-lingual deception classification. Finally, we propose to better understand multi-modal deception detection and explore methods to weight and combine information from multiple modalities to improve multi-modal deception classification.

1 Introduction

Deception detection is a deliberate choice to mislead to gain some advantage or avoid some penalty (DePaulo et al., 2003). Deception detection is an important goal of law enforcement, military and intelligence agencies, as well as commercial organizations. In recent years, automatic deception detection in text has gained popularity in the Natural Language Processing (NLP) community, and researchers have studied cues to deception in a diverse set of domains. These include detecting deception in news (Wang, 2017), online reviews (Ott et al., 2011), interview dialogues (Levitan et al., 2018), trial testimonies (Fornaciari and Poesio, 2013), and in games (Soldner et al., 2019). These studies have been useful for identifying linguistic characteristics of deception, and for developing machine learning techniques to automatically detect deceptive language. However, we are still a long way from applying these state-of-the-art deception detection models in real-world deception scenarios. We currently lack information about how deception detection models perform across domains, languages, and in multiple modalities. In this proposal we outline current limitations in these three areas of deception detection: across domains, across languages and in multiple modalities. We propose work to address these limitations, with the goal of developing robust deception detection models that can generalize from lab-based datasets to real-world deception.

For each of the three topics of deception detection, we discuss current limitations, formulate research questions, and state proposed work to address the research questions. For some of our research questions, we present completed or ongoing work to answer the questions. For cross-domain deception classification, we first establish baseline performance at within and cross-domain deception classification using the well-established NLP model BERT. We identify major performance gaps between within and cross-domain deception classification. To understand the cross-domain performances, we formulate distance metrics and propose a cross-domain classification model that does not require target domain labeled training data and outperforms several baseline models. We also discuss ongoing and future research to further our understanding of and further imrpove cross-domain deception detection.

For cross-lingual deception classification, we formulate the task for deception detection in two non-English languages: Bulgarian and Arabic. We discuss the effectiveness of using a wide range of classifiers including multilingual BERT (Devlin et al., 2019), and propose additional experiments to further understand and improve cross-lingual deception detection.

Finally, we present proposed work in deception detection in a multi-modal setting from text and image features. Learning to identify deception is a challenging task, especially when there is one modality, and we propose to dynamically fuse in-

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Student Research Workshop, pages 383 - 390
May 22-27, 2022 ©2022 Association for Computational Linguistics

formation from multiple modalities. The thorough experiments in the proposed work will contribute substantially to our understanding of cross-domain, cross-lingual, and multi-modal deception detection, and to the development of robust deception detection technologies.

2 Current limitations

2.1 Cross-domain deception classification

Although deception detection is a popular task in the NLP research community, and there is a strong interest in commercial applications of this work, there exists a large gap between deception models trained under laboratory conditions, and the performance level that is needed in real-world deception. Although researchers have in some cases obtained very strong performance at deception detection, these studies have focused on single domains, often using small datasets. We currently lack information about how small-scale, single-domain models of deception may or may not generalize to real-world data and new domains. We directly address this gap by first benchmarking the within- and cross-domain deception classification performance using five popular deceptive text datasets. We then attempt to understand performance gaps by analyzing the features of the datasets and the learned embeddings representations by the models. Finally, we propose a novel approach to leverage distance between domains to improve cross-domain deception classification.

Studying cross-domain deception detection is critical for understanding and contextualizing the successes of deception detection models thus far and gaining insights about the unique challenges of deception detection. The insights gained will motivate and inform the development of more robust models of deception.

2.2 Cross-lingual deception classification

There has been recent work in the NLP community aimed at identifying general misinformation on social media (Shu et al., 2017; Mitra et al., 2017) and particularly COVID-19 misinformation (Hossain et al., 2020).

However, most of this prior work has focused on data in English. There is a severe data shortage of high quality datasets that are labeled for misinformation in multiple languages. Because of this, we need to develop models of deception and misinformation that can utilize smaller datasets in non-English languages or leverage large amounts of training data in a source language, such as English, and generalize to new target languages.

2.3 Multi-modal deception classification

To build classifiers that can detect deception at a high accuracy, it is necessary to have high quality training data. Although a lot of prior work has focused on predicting deception from text (Potthast et al., 2017; Ott et al., 2011; Fornaciari and Poesio, 2014; Levitan et al., 2018), it is generally harder to identify deception from just one modality.

Nakamura et al. (2020) propose models to combine the image and text modalities by simple concatenation, addition, subtraction or taking dimension-wise maximum of image and text feature vectors. However, it still seems unclear how much importance should be attributed to each modality. Whether there are better ways to combine modalities is still unknown.

3 Proposed work and preliminary exploration

To address the limitations discussed above, we formulate concrete research questions that would help study deception classification across domains, languages and modalities. We now discuss proposed work and findings from initial experiments for each research question.

3.1 Cross-domain deception classification

RQ1. How do current models of deception perform within domain and across domain?

To address this research question, we select five deception datasets from different domains for our analysis. They were selected because they are all publicly available, and have been widely used for training and evaluating within-domain deception detection performance. This collection of datasets includes (1) *Fake news* containing fake and legitimate news compiled via a combination of crowdsourcing and webscraping (Pérez-Rosas et al., 2018), (2) *Open-domain deception* consisting of short, open-domain truths and lies obtained via crowdsourcing (Pérez-Rosas and Mihalcea, 2015), (3) *Cross-cultural deception* consisting of a set of deceptive and truthful essays about three topics: opinions on abortion, opinions on death penalty, and feelings about a best friend (Pérez-Rosas and Mihalcea, 2014), (4) *Deceptive opinion spam* containing truthful and deceptive hotel reviews of 20

Domain	Deception type	Number of tokens				Number of samples			
		Mean	Std.	1%ile	99%ile	Truthful	Deceptive	Train	Test
FakeNews	Self reported	324.50	692.35	78.58	1936.71	490	490	784	196
OpenDomain	Self reported	10.59	5.19	5.00	31.00	3584	3584	5734	1434
CrossCultural	Self reported	81.47	32.06	24.99	177.04	200	200	320	80
DeceptiveOpinion	Self reported	167.79	98.93	40.99	504.00	800	800	1280	320
Liar	Obs. reported	20.21	11.46	6.00	46.00	4507	8284	10232	2559

Table 1: Summary statistics for datasets from different domains along with distribution of truthful and deceptive classes and train/test sizes.

Chicago hotels (Ott et al., 2011), and (5) *Liar liar pants on fire* containing a set of short statements, mostly by politicians, in various contexts spanning across a decade (Wang, 2017). Since each dataset was collected under different experimental settings and have different topics and styles, we consider each dataset to represent a different domain without loss of generality. The summary statistics of the datasets in each domain are shown in Table 1. 4 of the 5 datasets have perfectly balanced classes, while *Liar* has approximately 35% truthful samples and 65% deceptive samples. It is important to note that the method of obtaining deception labels can vary for different datasets. Broadly, each dataset can be categorized into self reported deception or observed reported deception, based on whether they were reported by the speakers/writers or by human labelers respectively. We show the deception type for various datasets in Table 1. We perform a stratified splitting of the dataset of each domain into training and test splits with 80% of the data used for training and 20% used for testing, sizes of which are shown in Table 1. These train/test splits are used consistently across all experiments in this work to ensure a fair comparison of results across experiments.

We applied a state-of-the-art NLP model BERT (Devlin et al., 2019) to establish a strong baseline for cross-domain deception detection. We used a 10% random split of the source domain training data as the development data. For deception classification, we fine-tuned a BERT-based sequence classification model.[1] For training the BERT-based model the Adam optimizer (Kingma and Ba, 2014) was used with a learning rate of 1e-5. The training was stopped when the development accuracy did not improve for 5 consecutive epochs.

We observe in Table 2 that for any given target domain, the in-domain accuracies are generally higher than the cross-domain accuracies. This finding is consistent with observations made by Glenski et al. (2020). In some cases, the gap between within and across domain performance is egregious. For example, BERT classifier fine-tuned on *DeceptiveOpinion* has a within domain accuracy of 0.909, while the cross-domain performance of a model trained on *DeceptiveOpinion* ranges from 0.453-0.550 for the four other target domains. Further, the cross-domain performance of models trained on other domains and tested on *DeceptiveOpinion* ranges from 0.456-0.572. Although the *DeceptiveOpinion* model has very strong within domain performance and is a useful model of deceptive hotel reviews, it is clearly not a robust model of deception and cannot generalize to other deception domains.

RQ2. When there is a performance gap between within and across domain deception detection, can we explain why that occurs?

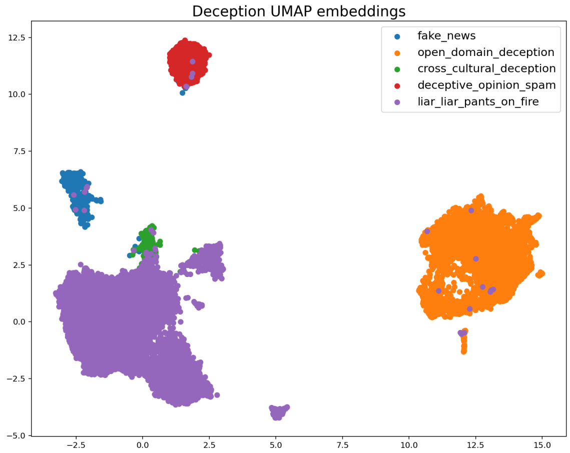

Figure 1: Deception sentence embeddings from different domains using pre-trained BERT.

To gain a deeper understanding of the classification results, we take BERT `[CLS]` token's representation to extract sentence level embedding of each sentence. To visualize the deception sentence embeddings, we project the sentence embeddings into a 2D space using UMAP (McInnes et al.,

[1] `bert-base-uncased` model in Transformers library (Wolf et al., 2020).

Target domain →	***FakeNews***	***OpenDomain***	***CrossCultural***	***DeceptiveOpinion***	***Liar***
FakeNews	0.786	**0.518**	0.5	**0.572**	**0.62**
OpenDomain	0.474	0.642	0.4	0.478	0.581
CrossCultural	**0.566**	0.504	0.613	0.456	0.501
DeceptiveOpinion	0.52	0.5	**0.55**	0.909	0.453
Liar	0.5	0.504	0.5	0.506	0.674

Table 2: In-domain and cross-domain accuracies of deception detection. For each target domain, the in-domain accuracy is underlined and the best cross-domain accuracy is bold-faced.

2018). We observe from Figure 1 that there are well-defined clusters of embeddings for most domains, for example *DeceptiveOpinion*, in red). In contrast, the *Liar* dataset, shown in purple, appears to have more broad and diverse embeddings, with several purple data points appearing in each of the other clusters.

We analyze the sentence embeddings further by defining a distance metric which can be used to measure the distance between a pair of domains. We first formulate the general notion of distance. Let D_S and D_T be the source and target domains respectively. We denote the distance from D_S to D_T as $distance(D_S, D_T)$. The distance between D_S and D_T can be computed using sentence embeddings as

$$distance(D_S, D_T) = \frac{1 - \cos(SD_S, SD_T)}{2}, \tag{1}$$

where SD_S is the mean of all the sentence embeddings in D_S and SD_T is the mean of all the sentence embeddings in D_T. Upon computing the Pearson correlation between cross-domain distances and accuracies, the correlation coefficient is found out to be -0.519, asserting that $distance(D_S, D_T)$ is negatively correlated with the cross-domain accuracy as expected.

We propose to understand the cross-domain deception performance by analyzing linguistic features of text in different domains. The list of linguistic features include politeness (Danescu-Niculescu-Mizil et al., 2013), concreteness (Kleinberg et al., 2019), complexity (Lu and Ai, 2015), readability (Dubay, 2004), sentiment and lexical features such as sentence length. Our analysis will include finding out the linguistic features correlated to deception for each domain and comparison of these features across domains.

RQ3. Can we leverage our understanding of these performance gaps to improve cross-domain deception detection?

We aim to develop a classification approach that leverages the notion of domain distance to improve cross-domain deception detection. The main idea is as follows: given a target domain, find the optimal source domain to use for training a deception detection model. We compute the domain distance between the target domain and all possible source domains. Then, we recommend the source domain which has the smallest distance from the target domain.

We compare the performance of this recommender system with 2 baselines: (1) A random recommendation system which chooses a source domain uniformly at random for a given target domain. To get a reliable cross-domain accuracy, we consider 100,000 trials of random recommendation and calculate the average cross-domain accuracy across all trials. (2) Multi-source leave-one-out training, which combines all source domains, excluding the target domain, for classification. The recommendation results are shown in Table 3. The table shows the accuracy upon using the recommended source domain for a given target domain. We observe that the recommendation using sentence embeddings based distance metrics is better than both random recommendation and leave one out multisource recommendation. This is an important use case of distance metrics, showing that they can reliably be used for improving cross-domain performance.

We find in Table 3 that while recommending a source domain is a relatively easier task for some target domains, recommendation is difficult in some other domains. For example, for the target domains *FakeNews* and *OpenDomain*, the recommendation using average sentence embeddings is right in a majority of cases. However, this is more challenging for *Liar* as the target domain, since no model achieves an accuracy that is substantially above 50%. To improve recommendation for such cases, we propose to compute the distance between a sample and all the potential source domains using sentence embeddings. By doing the recommendation at a sample level, we hope to improve the overall prediction on the target domain.

Recommendation	Target domain				
	FakeNews	*OpenDomain*	*CrossCultural*	*DeceptiveOpinion*	*Liar*
Random recommendation	0.553	0.484	0.507	0.506	0.503
Multisource leave one out	0.541	0.500	**0.550**	0.447	**0.521**
Avg. sentence embed.	**0.620**	**0.581**	0.501	**0.550**	0.500
Best possible recommendation	0.620	0.581	0.566	0.550	0.506

Table 3: Cross-domain accuracies upon recommending for various target domains.

Language →	English	Bulgarian	Arabic
Train	869	3000	2536
Dev	53	350	520
Test	418	357	1000
Total	1340	3707	4056

Table 4: Data sizes of English, Bulgarian and Arabic datasets for COVID-19 misinformation detection.

Setup	Eng. → Bulgarian	Eng. → Arabic
Zero shot	0.810	0.672
Few-50	0.819	0.775
Few-100	0.823	0.824
Few-150	0.821	0.791
Full shot	0.834	0.787
Target	0.843	0.738

Table 5: Cross-lingual (source language → target language) F1 scores when tested on the target language. *Few-n* setup denotes that only n samples in the target language are used for training.

3.2 Cross-lingual deception classification

RQ4. How effective are state of the art multilingual NLP models at cross lingual deception classification?

To answer this question, we use the findings from Panda and Levitan (2021) who used the tweet data provided for the Fighting the COVID-19 Infodemic shared task (Shaar et al., 2021) for analysis. The data was created by answering 7 questions about COVID-19 for each tweet about the following aspects: verifiable factual claim, false information, interest to general public, harmfulness, need of verification, harmful to society, and require attention. Each question has a Yes/No (binary) annotation. The data includes tweets in three languages: English, Bulgarian and Arabic. The data falls in the observed reported deception category (see the data discussion in Section 3.1). The training, development and test data sizes for each of the three languages are shown in Table 4. An example of an English tweet from the dataset is *Anyone else notice that COVID-19 seemed to pop up almost immediately after impeachment failed?* The 7 corresponding labels are *Q1 Yes, Q2 Yes, Q3 Yes, Q4 Yes, Q5 No, Q6 Yes, Q7 No.*

When the features from multilingual BERT (Devlin et al., 2019) are used for training on the source language and then testing is done on the target language, the scores as reported in Panda and Levitan (2021) are shown in Table 5. This is the *Zero shot* setup. The source language is set to English and the target languages are Bulgarian and Arabic. The scores for training using the target language (*Target* setup) are also shown for comparison. We observe that the cross-lingual F1 scores in the *Zero shot* setup are lower than the scores in the *Target* setup. Without the target language training data, the model as expected finds it harder to predict accurately when tested on the target language.

RQ5. What is the impact of amount of target language training data on prediction quality?

To answer this question, all the source language training data combined with n training samples from the target language is used for training. n is to 50, 100 and 150. This is called the *Few shot* setup. A special case of this setup is the *Full shot* setup, where n is set to the total size of the target language training data. We observe in Table 5 that as we increase the target language training samples in the few shot setup, the performance increases in general, as one would expect. Notably, even with just 50 samples from the target language training data, there is a noticeable increase in the cross-lingual performance in comparison to the *Zero shot* setup.

RQ6. How effective is using machine translation for cross lingual deception classification?

We propose to study the effectiveness of translation with multilingual COVID-19 misinformation classification models. In most cases, training data is available in English. The main idea is to translate either the training or the test non-English data to English using a pre-trained machine translation system. We plan to use the state-of-the-art machine translation systems by Tiedemann and Thottingal (2020) to

1. Translate the non-English test set to English and use an English model for prediction.

2-way classification			
Category	**Train**	**Dev**	**Test**
True	215490	22585	22798
Fake	337449	35567	35309
— Total —	552939	58152	58107
3-way classification			
Category	**Train**	**Dev**	**Test**
Completely True	215490	22585	22798
Fake with False text	323721	34217	33835
Fake with True text	13728	1350	1474
— Total —	552939	58152	58107
6-way classification			
Category	**Train**	**Dev**	**Test**
True	215490	22585	22798
Misleading content	104136	10970	10959
False connection	167471	17766	17429
Manipulated content	21437	2161	2286
Satire/parody	32718	3438	3419
Imposter content	11687	1232	1216
— Total —	552939	58152	58107

Table 6: Data sizes for different categories for multi-modal deception classification.

2. Translate English training data to a target language and train the m-BERT classification model using this translated data.

Results from the above experiments will help quantify the effectiveness of using translation for deception detection.

3.3 Multi-modal deception classification

RQ7. Are there better ways to fuse text and image features in comparison to static fusion?

Recent work such as Nakamura et al. (2020) have created multimodal deception datasets and also provided baselines of fusing multiple modalities. The dataset by Nakamura et al. (2020), called Fakeddit, is the largest publicly available multi-modal deception dataset. It contains two modalities: text and image. There are three labels for each data sample, varying on granularity. The fine-grained labels are true content, misleading content, false connection, manipulated content, satire/parody and imposter content. These labels come from subreddits in from which the content is taken from (see details in Nakamura et al. (2020)). This dataset can be used to train a classifier to predict deception on a desired fine-grained level, the choices of which are 2-way, 3-way and 6-way. The Fakeddit dataset falls under the self reported deception category (see the data discussion in Section 3.1). The sizes of the Fakeddit dataset are shown in Table 6.

We propose to use an attention module (Luong et al., 2015) that dynamically fuses the text and image feature vectors as shown in Figure 2. The text feature vector comes from the `[CLS]` token's representation of BERT. The image feature vector comes from ResNet. The attention module decides how much weight to assign for each modality. Specifically it uses the input features for each modality and computes as many attention scores as the number of modalities. These attention scores are positive and sum to 1 when added together. The feature vector from each modality is scaled using the corresponding attention score. Then the scaled feature vectors are added together to obtain a single vector, which can be passed through a final linear layer to produce logits. We plan to answer the following questions by analyzing the results of attention-based fusion.

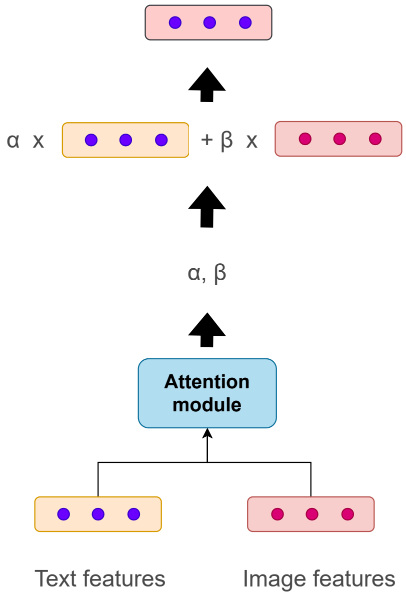

Figure 2: Fusion of text and image features using an attention module.

1. For each category of samples, what is the average attention on each modality?

2. Are there samples for which the attention to one modality is negligible? Are there patterns among these samples?

3. Does dynamic fusion of the text and image feature vectors lead to better overall prediction than static fusion?

4 Ethical considerations

Although automatic deception detection has the potential to benefit society, there are several ethical concerns within this line of research. Automatic deception detection has varying degrees of severity depending on the application area. The impact of a false positive is substantially lower when detecting deceit in informal activities such as gaming. However, when detecting dishonesty in a criminal investigation, a false positive can have serious implications. In general, automatic deception detection should be employed with caution, especially when there is no manual human verification involved.

For the case of cross-domain deception detection applications, it is important to test the model on the target domain before deploying it, as mentioned in Section 1. To understand the differences between deception domains, a linguistic feature analysis should be performed, as we mention in Section 3.1. Finally, to increase transparency in multi-modal deception detection, it is critical to compute importance scores for each modality as mentioned in Section 3.3. As automatic deception detection across domains, languages and modalities becomes a more widely studied subject, it is important to be aware of the ethical considerations and also take the necessary precautions to avoid harm to society.

5 Conclusion

We identify key challenges in deception detection in cross-domain, cross-lingual and multi-modal scenarios. For cross-domain deception classification, we quantified the gap between in-domain and cross-domain accuracies. Our proposed recommender based on distance measures improves cross-domain performance over two baselines. We plan to extend the completed work by improving the recommendation process by recommending at the sample level instead of the domain level. We also plan to analyze the cross-domain results using linguistic features. For cross-lingual deception classification, we discuss the challenges in predicting deception in a target language with no or little training data. We propose to study the effectiveness of using translation text for training and testing. For multi-modal deception classification, we discuss the merits and limitations of the current state-of-the-art models. We propose to dynamically fuse the text and image feature vectors using an attention module to better understand the importance of each modality.

References

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 250–259, Sofia, Bulgaria. Association for Computational Linguistics.

Bella DePaulo, James J Lindsay, Brian Malone, Laura Muhlenbruck, Kelly Charlton, and Harris Cooper. 2003. Cues to deception. *Psychological bulletin*, 129:74–118.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

William Dubay. 2004. The principles of readability. *CA*, 92627949:631–3309.

Tommaso Fornaciari and Massimo Poesio. 2013. Automatic deception detection in italian court cases. *Artificial intelligence and law*, 21(3):303–340.

Tommaso Fornaciari and Massimo Poesio. 2014. Identifying fake amazon reviews as learning from crowds.

Maria Glenski, Ellyn Ayton, Robin Cosbey, Dustin Arendt, and Svitlana Volkova. 2020. Towards trustworthy deception detection: Benchmarking model robustness across domains, modalities, and languages. In *Proceedings of the 3rd International Workshop on Rumours and Deception in Social Media (RDSM)*, pages 1–13, Barcelona, Spain (Online). Association for Computational Linguistics.

Tamanna Hossain, Robert L Logan IV, Arjuna Ugarte, Yoshitomo Matsubara, Sean Young, and Sameer Singh. 2020. Covidlies: Detecting covid-19 misinformation on social media. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. Cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

Bennett Kleinberg, Isabelle van der Vegt, Arnoud Arntz, and Bruno Verschuere. 2019. Detecting deceptive communication through linguistic concreteness.

Sarah Ita Levitan, Angel Maredia, and Julia Hirschberg. 2018. Linguistic cues to deception and perceived deception in interview dialogues. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1941–1950.

Xiaofei Lu and Haiyang Ai. 2015. Syntactic complexity in college-level english writing: Differences among writers with diverse l1 backgrounds. *Journal of Second Language Writing*, 29:16–27. New developments in the study of L2 writing complexity.

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

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. 2018. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861.

Tanushree Mitra, Graham P Wright, and Eric Gilbert. 2017. A parsimonious language model of social media credibility across disparate events. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, pages 126–145.

Kai Nakamura, Sharon Levy, and William Yang Wang. 2020. Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6149–6157, Marseille, France. European Language Resources Association.

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*, pages 309–319, Portland, Oregon, USA. Association for Computational Linguistics.

Subhadarshi Panda and Sarah Ita Levitan. 2021. Detecting multilingual COVID-19 misinformation on social media via contextualized embeddings. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 125–129, Online. Association for Computational Linguistics.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Verónica Pérez-Rosas and Rada Mihalcea. 2014. Cross-cultural deception detection. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 440–445, Baltimore, Maryland. Association for Computational Linguistics.

Verónica Pérez-Rosas and Rada Mihalcea. 2015. Experiments in open domain deception detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1120–1125, Lisbon, Portugal. Association for Computational Linguistics.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2017. A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint arXiv:1702.05638*.

Shaden Shaar, Firoj Alam, Giovanni Da San Martino, Alex Nikolov, Wajdi Zaghouani, Preslav Nakov, and Anna Feldman. 2021. Findings of the NLP4IF-2021 shared task on fighting the COVID-19 infodemic and censorship detection. In *Proceedings of the Fourth Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, NLP4IF@NAACL' 21, Online. Association for Computational Linguistics.

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.

Felix Soldner, Verónica Pérez-Rosas, and Rada Mihalcea. 2019. Box of lies: Multimodal deception detection in dialogues. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1768–1777.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.