HOW GENERALIZABLE ARE DEEP CONTEXTUAL MODELS?

Mohammad Rashedul Hasan

Department of Electrical and Computer Engineering, University of Nebraska-Lincoln, Lincoln, Nebraska, USA

ABSTRACT

We investigate the generalizability of the deep contextual models along two dimensions: (i) when data includes unreliable or noisy categories and (ii) when data is out-of-distribution (OOD). Specifically, we focus on the Transformer-based BERT (Bidirectional Encoder Representation from Transformer) model for recognizing COVID-19 misinformation data from online social media. A set of studies are designed to examine the generalizability of a diverse array of BERT-based transfer learning techniques. The investigation also includes shallow non-contextual models. Results obtained from extensive systematic experimentation show that the BERT-based models generalize poorly on the OOD data as well as when the domain contains unverified samples. Notably, these deep contextual models are not more effective, and at times worse, than shallow non-contextual models.

We explain possible reasons for the poor generalizability of deep contextual models.

KEYWORDS

Deep Contextual Models, Generalizability, Natural Language Processing, COVID-19, Noisy Data

1. INTRODUCTION

Though Transformer-based [1] deep contextual models have achieved state-of-the-art performance on some static benchmark Natural Language Processing (NLP) datasets [2, 3], it is not clear yet how generalizable these models are when used in practical datasets that are noisy and dynamic [4]. The challenge of generalizability arises from the nature of both training and test data. When training data is contaminated with noisy labels, deep learning models' generalizability degrades substantially [5]. Generalizability also stumbles when a model is applied to out-of-distribution (OOD) data [6].

A common-sense heuristic is that to be generalizable, deep contextual models must acquire a deep understanding of the language [4]. Transformer-based models aspire to achieve this "understanding" by learning language representations from a general-purpose unlabeled source data that is amenable to a downstream task (e.g., text classification) via transfer learning [7]. These representations capture semantic and syntactic relationships of the words (i.e., complex characteristics of word use) as well as their contextual relationships (i.e., polysemy) [8]. However, these properties of a deep contextual pretrained model (PTM) may not be enough to ensure its generalizability. For example, if the target domain contains samples with unreliable or noisy labels, then transfer learning may not yield an optimal performance. Due to the involvement of non-expert labelers [9] as well as when the expert-labelers lack the domain knowledge [10], existence of noisy labels in practical datasets is unavoidable. On the other hand,

David C. Wyld et al. (Eds): SIGI, CSTY, AI, NMOCT, BIOS, AIMLNET, MaVaS, BINLP – 2025
pp. 179-196, 2025. CS & IT - CSCP 2025 DOI: 10.5121/csit.2025.151915

even if the training data has reliable labels, a shift in the data distribution in practical problems results in poor generalization. This issue becomes severe when the domain is expected to go through a continuous shift in distribution. For example, in the domain of online social media, the language changes continuously [11], as a result, the test data distribution drifts from the train distribution over time [12]. To cope well with this type of OOD data, NLP techniques are not only required to learn the language model but also to understand the changing pattern in the language [13]. Since the development and evaluation of the deep contextual models occur under the assumption that train and test samples are independent and identically distributed (i.i.d.) [14], it is not apparent whether these models understand the language change for being able to generalize over OOD data.

In this paper, we systematically study the generalizability of the deep contextual model BERT (Bidirectional Encoder Representation from Transformer) [15]. Specifically, we investigate whether BERT-based models are generalizable when (i) training data samples contain noisy labels, and (ii) test data is OOD. As an example of an NLP task that captures these two challenges of generalizability, we focus on the problem of misinformation detection from online social media data related to Coronavirus or COVID-19 pandemic. Designing an effective text classifier for this problem is a daunting task due to the nature of the COVID-19 social media data. Unlike the curated static NLP datasets on which deep contextual models like BERT are tested [16], COVID-19 online misinformation datasets are noisy and dynamic. Creating reliable labels for COVID-19 misinformation data is both an expensive and a time-consuming task. As a consequence, not only misinformation datasets may contain samples with noisy labels [17], but also there may exist noisy categories (e.g., an entire category could be labeled as "unverified" due to lack of knowledge during the time of data collection [18]). The dynamic nature of COVID-19 misinformation is due to the variation in the misinformation narrative across geographic regions [19] as well as variation over time (caused by the faster evolution of misinformation themes [20]). As a result of these two dynamic aspects of the COVID-19 data, a contextual model developed using localized data or data collected from a specific duration of time, may find it challenging to generalize over data from different regions or future periods, which are OOD. Previously deep contextual models including BERT were utilized to design text classifiers for detecting COVID-19 misinformation [18, 21, 22, 23]. However, the train and test samples used for the development of these approaches were i.i.d., i.e., test data is randomly selected from the dataset used for training. As a consequence, the generalizability of these approaches has not been verified yet.

Our generalizability study on the deep contextual model BERT spans along two dimensions: we use (i) train data with unreliable categories to evaluate models on in-distribution test data, and (ii) train data with reliable categories to evaluate models on OOD test data. Moreover, we include a diverse set of OOD data with varying degrees of distribution shift. For a comparative understanding, our study includes shallow non-contextual models such as Word2Vec [24] and FastText [25], which are based on shallow neural networks.

We examine a diverse set of techniques, both contextual and non-contextual, for the study. The techniques are broadly divided into two paradigms of transfer learning: (i) it involves domainagnostic (DA) pretrained models (PTM) that learns language representations from general-purpose unlabeled data [7], and (ii) it involves domain-specific (DS) PTMs that learns representations from domains that are similar to the target domain [26]. Two knowledge-transfer approaches (both for DA and DS BERT PTMs) are used [27]: (i) extracted feature-based (FB) learning in which the BERT-extracted features such as word embeddings from the model's output are fed into another neural network for training using the target data, and (ii) adding a classification layer on top of the BERT PTM, then fine-tuning (FT) its hidden layers using the target data. The non-contextual Word2Vec and FastText models are used only as feature

extractors in FB learning (both as DA and DS PTMs). These models are pretrained to learn representations of a set of words from a source dataset. The representations or word embeddingsare then transferred as features to train a text classifier by using the target data [28].

Contributions. We design a set of studies to conduct a deeper investigation of the generalizability of the deep contextual BERT model. We emphasize two dimensions of generalizability (in presence of noisy categories and varying-degree of OOD data) that has not yet been explored. Our main contributions include the following findings.

- The deep contextual BERT-based models (both DA and DS) do not generalize well (i) in presence of unreliable categories in the training data, and (ii) on OOD data.
- Unlike the observed superior generalizability of contextual models in [14], we find that the shallow non-contextual Word2Vec and FastText-based models exhibit competitive and sometimes better performance over BERT.
- We show that there could be considerable variation in the distribution shift across OOD datasets. We examine the models at the backdrop of the varying space of OOD. We show that depending on the nature of the OOD data (e.g., spatially-varying or temporally-varying) and the degree of the distribution shift, the generalizability performance of the models varies.
- We explain the lack of robustness of the deep contextual models. While the BERT-based models can learn contextual representations within the static space of the source data, they are not good at understanding the language change. Even when a BERT PTM is created by using COVID Twitter data (e.g, the COVID-Twitter-BERT (CT-BERT) [29]), it does not generalize well on the COVID-19 OOD data due to its lack of understanding of the language change. The priors (learned from the source data) of the BERT PTMs are much stronger than those of the non-contextual models, which may have imposed heavy inertia on their adaptation capability in the latent space for capturing the language change present in the target data.

2. RELATED WORK

In the machine learning based NLP, the text input data is encoded with latent representations or embeddings amenable for solving a downstream task. These embeddings are learned by neural pretrained models (PTMs) from general-purpose unlabeled data by using the self-supervised learning approach [30]. Then, the embeddingsare transferred to the downstream task either via fine-tuning or by feature extraction [27].

Shallow Non-contextual Models. The PTM Word2Vec [24] and FastText [25] are predictive models that are based on shallow neural networks. Both models learn word embeddings from the unlabelled Wikipedia corpus. Word2Vec uses two types of models for learning: Continuous Bag-of-Words (CBOW) and Skip-Gram. CBOW learns embeddings by predicting the most likely word in the given context, while in Skip-Gram the model learns by predicting the context using the given word. Both Word2Vec and FastText learn non-contextual word embeddings from their co-occurrence information. The main limitation of Word2Vec is that it is unable to encode out-of-vocabulary words. FastText overcomes this limitation by extending the Word2Vec model. Specifically, it first breaks the words into several sub-words (or n-grams) and then feeds them into the neural network.

Deep Contextual Models. Unlike Word2Vec and FastText PTMs, the BERT PTM [15] can learn contextual embeddings. It utilizes an autoencoding technique with bi-directional context modeling. BERT is based on the Transformer model [1], which is a very deep neural

architecture equipped with a multi-head attention mechanism. Two variants of the BERT architecture are generally used: BERT Large and BERT Base. BERT Large uses 24 encoder layers with 24 bidirectional self-attention heads each with 1042 hidden dimensions, while BERT Base uses 12 encoder layers with 12 bi-directional self-attention heads each with 768 hidden dimensions. Both variants are pretrained using unlabeled data extracted from the BooksCorpus with 800M words and English Wikipedia with 2,500M words. BERT learns the embeddings by utilizing the surrounding context signals from the text corpora. Specifically, it learns using two predicting tasks: by predicting random missing words (15%) using the rest of the sentence (i.e., a masked language modeling task); and by predicting whether two sentences appear next to each other. BERT provides a [CLS] token at the start of the sequence, whose embeddingsare treated as the representation of the text sequence(s).

3. DATASET, STUDY DESIGN AND MODELS

In this section, we describe the dataset pre-processing, design of the studies, and the DA as well as DS PTM-based transfer learning approaches that include both shallow non-contextual and deep contextual models.

3.1. Dataset

We use a COVID-19 social media misinformation dataset collected by Princeton University's Empirical Studies of Conflict Project (ESOC) [19]. The dataset contains 5,613 distinct misinformation stories originated from social media posts such as tweets and news articles during the full year of 2020. These stories came from over 80 countries and spanned across 35 languages.

We have four reasons for choosing this dataset for the study of the generalizability of predictive models. (i) It is a multi-class misinformation dataset that contains tweets and news stories from social media belonging to three misinformation categories, i.e., false reporting, conspiracy, and fake remedy. It is challenging to design effective models by using only misinformation samples of various types. (ii) The dataset is heavily skewed having more than 75% of samples in the false reporting category. Thus, generalizing over the minority classes is a challenging task for the models. (iii) Though all misinformation samples belong to one of the three categories, there exists significant variation in the misinformation themes within the categories. The ESOC project report [19] shows that false narratives are localized, i.e., the nature of misinformation changes across regions and countries. This nuanced nature of misinformation makes it harder for the models to generalize. (iv) Finally, the dataset contains metadata that we leverage to create an array of diverse OOD test sets.

Pre-processing. For training and evaluating the models, we only used the text written in English. We extracted a total of 1,235 English text samples originated from predominantly Englishspeaking countries as well from countries where the primary language is not English, e.g., Hindi, Tagalog, Sinhala, Chinese, and Urdu. Out of the 1,235 samples, 951 samples belong to the false reporting category (class 0), 186 samples belong to the conspiracy category (class 1), and 98 samples belong to the fake remedy category (class 2). We consider these three categories reliable as they were labeled and verified by domain experts [19].

We use two metadata, i.e., primary language and publication date, to create two orthogonal OOD test sets. The first OOD test set contains misinformation samples that exhibit distribution shift along the dimension of geographic locations, while the second OOD test samples exhibit varying degrees of distribution shift along the temporal dimension. The process of creating these orthogonal OOD test sets is described next in the study design sub-section.

3.2. Study Design

Our generalizability study spans two dimensions.

- Dimension 1: Train data with unreliable categories and in-distribution test data.
- Dimension 2: Train data with reliable categories and OOD test data. Generalizability on OOD data is examined in two orthogonal dimensions.
  - Dimension 2(a): OOD data - shift along geographical dimension
  - Dimension 2(b): OOD data - shift along temporal dimension

First, we create a benchmark in study 1 by using train data with reliable categories and indistribution test data. In study 2, we evaluate generalizability of the models along dimension 1. Then, in studies 2 and 3, the models are evaluated on two orthogonal OOD data (dimensions 2(a) and 2(b)). Two additional studies 5 and 6 are designed to corroborate the observations of dimension 2 (studies 3 and 4). The last two studies are reported in the technical appendix.

Study 1 (Creating Benchmark): Train Data with Reliable Categories & In-Distribution Test Data: To create a benchmark, we use the train data with its three reliable categories and indistribution test data, i.e., train and test data are i.i.d. Specifically, we create train-test folds by randomly selecting 80% samples for training and 20% samples for testing.

Study 2 (Dimension 1): Train Data with an Unreliable Category & In-Distribution Test Data: We create an unreliable category by sampling 25% of the data from each of the three categories and labeling those sampled data with a new category called “unverified”. Thus, the new category contains noisy-labeled samples. Test data is i.i.d., created by sampling 20% of the data.

Study 3 (Dimension 2a): Train Data with Reliable Categories & OOD Test Data - Obtained From Disparate Geographic Locations: For this study, we divide the data based on whether it is originated from English or non-English speaking countries. The samples from English-speaking countries (a total of 1,042) are used for training and samples written in English by non-English countries (a total of 193) are used for testing. Since the COVID-19 misinformation themes are localized and vary across geographic regions [19], this test set can be considered as OOD.

Study 4 (Dimension 2b): Train Data with Reliable Categories & OOD Test Data - Obtained From Various Periods in Future: For creating train-test folds for this study, we consider the temporal dimension. By using the publication date metadata, we split all English samples into the following 5 subsets: January-April (658 samples), May-June (234 samples), July-August (151 samples), September-October (100 samples), and November-December (88 samples). The January-April subset is used for training and the remaining 4 subsets are used for testing. Since the models only see the January-April data, the unseen samples of the four test sets during MayDecember are from the “future”. Given the rapid propagation of COVID-19 data and dynamics in the nuanced narrative of misinformation [19, 31, 20], these four test sets from the “future” can be considered as OOD.

Sample distribution per class (both train and test) for all studies is given in the technical appendix.

3.3. Deep Contextual Model: BERT

The contextual BERT is used in both FT and FB learning. While FT is done using only the DA PTM (by following mixed-domain transfer learning protocol [32]), the FB learning utilizes both the DA as well as DS BERT PTMs.

DA BERT Fine-tuning (FT): For the FT experiments, we use the sequence classifier DA BERT PTM. This PTM adds a single linear layer on top of the BERT model. The pretrained weights of all hidden layers of the PTM and the randomly initialized weights of the top classification layer are adapted during FT. Two variants of DA BERT are used: BERT Base and BERT Large, obtained from the Hugging Face library [33]. These two variants are utilized to determine whether increased model capacity (i.e., BERT Large) improves generalization.

DA BERT Feature-based (FB) Learning: We use only the BERT Base model as a feature extractor. Two techniques are employed to extract the fixed embeddings, which are subsequently used to train a linear classifier. The first technique involves using the embeddings of the classification token (i.e., the [CLS] token), which is the first token of the last layer hidden state [34, 35]. These embeddingsare obtained by passing the target data through the BERT model. The second technique involves the extraction of the embeddings of the final hidden layer [36]. Then, global average pooling is applied for training a linear classifier.

DS BERT Feature-based (FB) Learning: We use the following DS BERT PTMs: SciBERT [26], Bio-Clinical BERT (BC BERT) [37], and COVID-Twitter-BERT (CT-BERT) [29]. These DS PTMs are chosen as their embeddings encode specifically the context of the health domain. The SciBERT model is pretrained using scientific papers from mostly the biomedical domain. The BC BERT model is trained on electronic health records from ICU patients at the Beth Israel Hospital in Boston, Massachusetts. These two BERT DS PTMs are based on the BERT Base model while the CT-BERT is based on the BERT Large model and is pretrained on a corpus of 160M tweets about the coronavirus during the period from January 12 to April 16 in 2020. These three models are obtained from the Hugging Face library [33]. We use the second feature extraction technique (presented in DA BERT FB Learning) for these models.

3.4. Shallow Non-Contextual Models: Word2Vec &FastText

The non-contextual models are used only as feature extractors (i.e., in FB learning). We utilize both the DA and DS embeddings from Word2Vec and FastText for transfer learning. In addition to this, we combine the DA and DS embeddings to see whether it improves generalization. The embeddings are used for the extraction of more expressive features via a Convolutional Neural Network (CNN), which is described at the end of this sub-section.

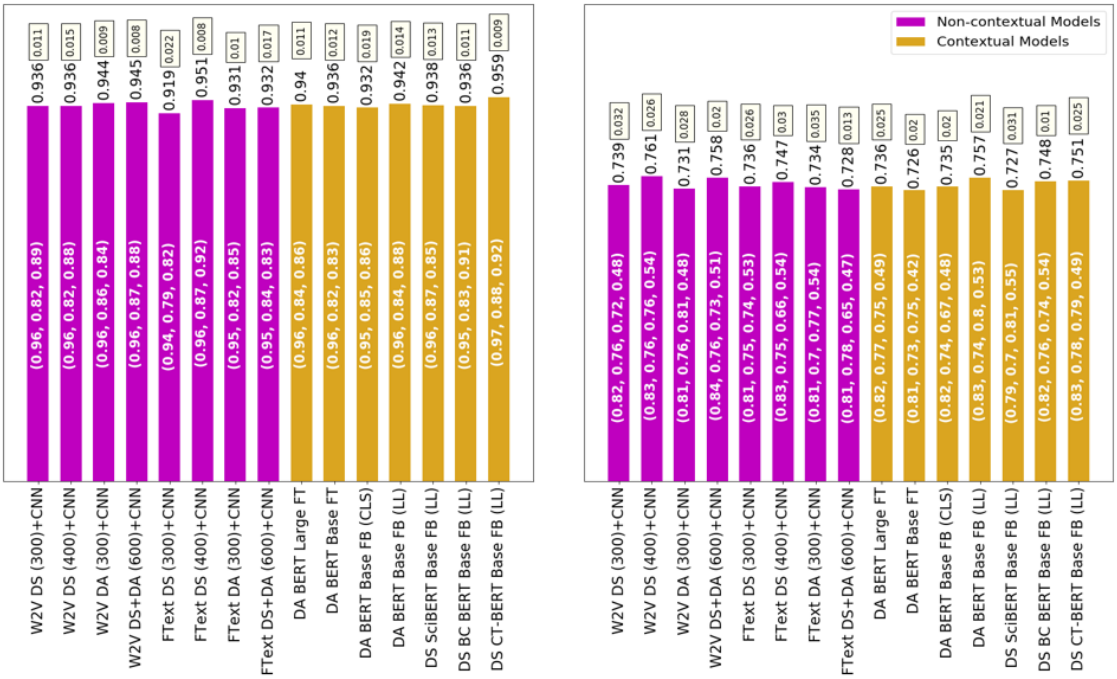
DA Embeddings for Feature-based (FB) Learning: The Word2Vec DA pretrained embeddings are obtained from Google Code [38]. The embedding vectors are 300-dimensional. We get the FastText DA pretrained 300-dimensional embeddings from [39].

DS Embeddings for Feature-based (FB) Learning: The DS embeddingsare learned by training the Word2Vec and FastText models using the target data. We create both 300-dimensional and 400-dimensional embeddings. Our goal is to see whether increasing the embedding dimension improves generalization. For creating the DS embeddings, we pre-process the data as follows. First, the text is converted to lower-case and tokenized, then single-character tokens are removed, followed by lemmatizing the tokens. Finally, the lemmatized tokens are used for learning their embeddings by the models.

Combine DS and DA Embeddings for Feature-based (FB) Learning: We concatenate the 300dimensional DA embeddings with the 300-dimensional DS embeddings and use the resulting 600dimensional embeddings for extracting higher-level features via a CNN.

Extraction of Expressive Features by a CNN: Both the DS and DA embeddings are used to train a CNN classifier that extracts higher-level and more expressive features by employing a single convolutional layer [28]. The CNN architecture consists of five layers. The first layer is the embedding layer. Its dimension varies based on the dimension of pretrained embeddings. The second layer is a one-dimensional convolution layer that has 200 filters of dimension 3 x 3 with "same" padding and ReLU activation. The third layer is a one-dimensional global max-pooling layer, and the fourth layer is a dense layer with 100 neurons along with ReLU activation. The last layer is the classification layer with softmax activation. We use this setting for the CNN architecture as it was found empirically optimal in our experiments. During the training, we adapted the DA, DS, and the concatenated word embeddings. Unlike in [28], we find the CNN classifier to be more effective when the embeddings are tuned.

4. Results and Analysis



(a) Study 1 (benchmark): Train data (reliable categories) & in-distribution test data. (b) Study 2: Train data (with an unreliable category) & in-distribution test data.

Figure 1: Avg. test accuracy (y-axis), standard deviation (top of each bar in boxes), and avg. F1 scores (shown inside the bars). "CLS": embeddings of the classification token; "LL": embeddings from the last hidden layer.

Experimental Setting. For learning DS Word2Vec and FastTextembeddings, we used the SkipGram model from the Gensim library [40]. For the validation purpose, 10% of the training data is used. The BERT-based models were trained for 10 epochs (both in FT and FB learning experiments) using the Rectified Adam optimizer [41] on a batch size of 16. The non-contextual embeddings based CNNs are trained for 20 epochs using the Adam optimizer on a batch size of

64. Each experiment was run 5 times, and an average accuracy, as well as standard deviation for the accuracy, are reported. In addition to this, the average F1 score for each class is presented. All experiments are done using Transformers, Scikit-learn, TensorFlow 2.0, and PyTorch libraries.

Study 1. Figure 1(a) shows the benchmark results from study 1. Since the train and test samples are i.i.d., *both non-contextual and contextual models generalize well, achieving above 90% test accuracy*. Though the training data was highly imbalanced (sample size for the 3 classes are 757, 159, & 72), all models obtain above 80% avg. F1 score for the smaller classes (i.e., classes 1 and 2). The CT-BERT FB and FastText DA (400) exhibit the best performance (above 95% avg. test accuracy and above 90% avg. F1 score on the smallest class 2). BERT Large FT performs slightly better than BERT Base FT. Also, the DS BERT models perform better than the DA BERT models.

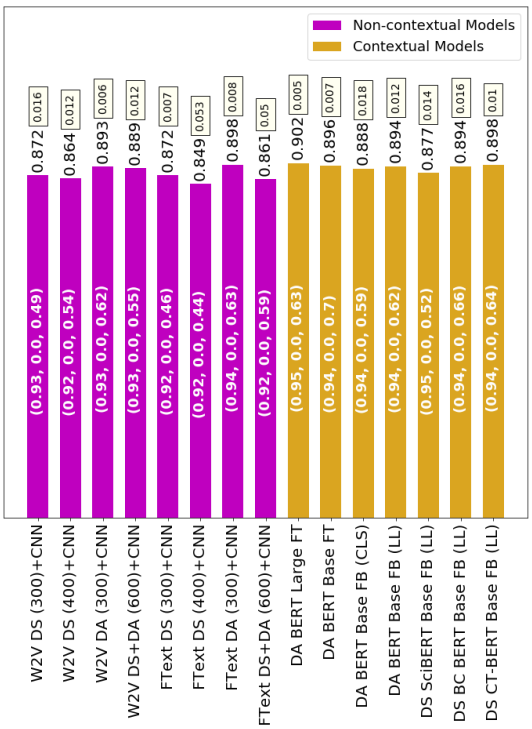
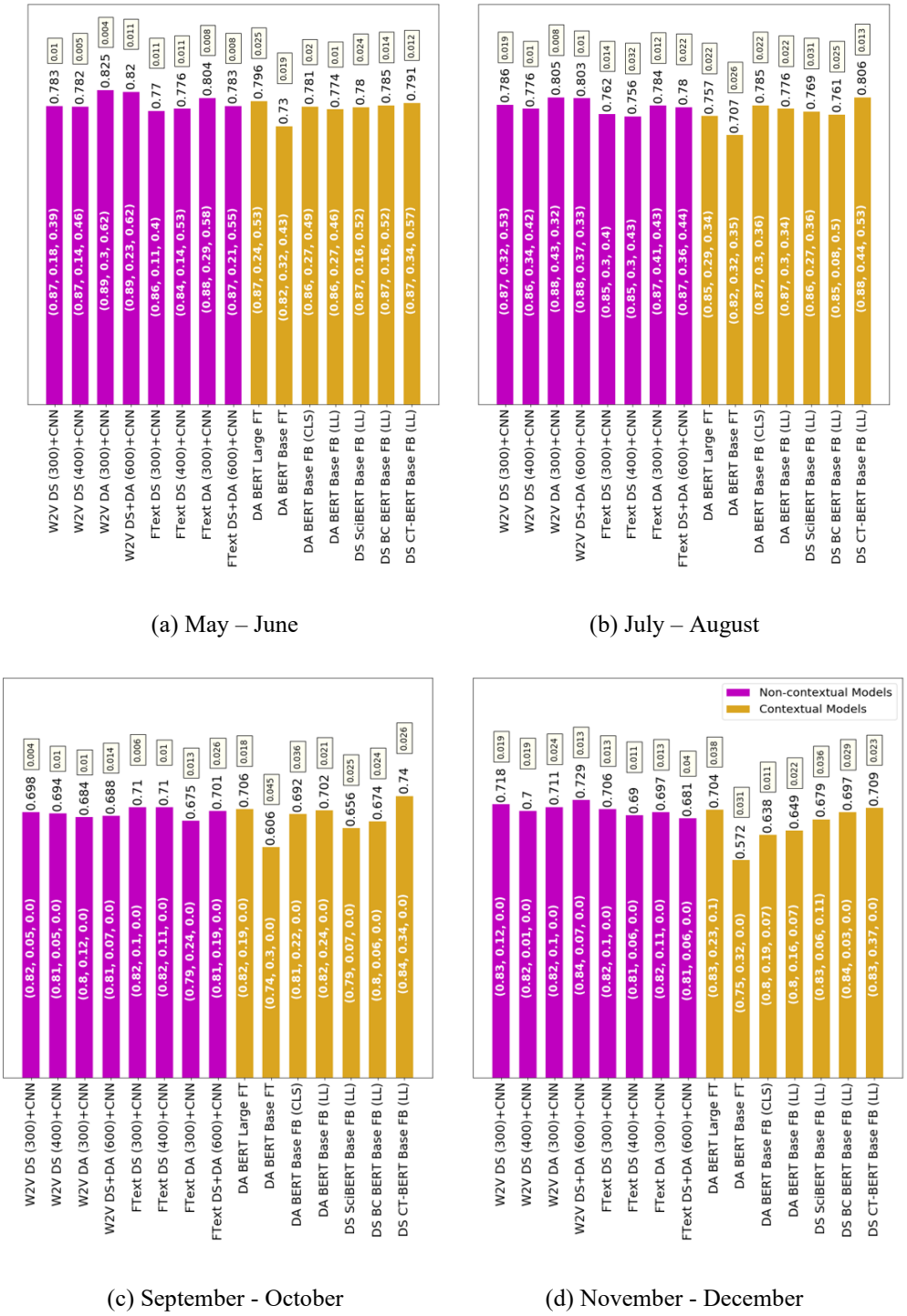


Figure 2: Study 3: Train data (reliable categories) = English speaking countries; Test data (OOD) = non-English speaking countries.

Study 2. The effect of the unreliable category is shown in Figure 1(b). Both the non-contextual and contextual models generalize poorly on in-distribution test data due to the presence of a noisy category. Interestingly, non-contextual Word2Vec performs slightly better than other models. Also, overall the FB BERT techniques exhibit better effectiveness than the FT-based techniques. Compared to study 1, there is an increase in the standard deviation for the test accuracies for most of the models.

Study 3. The generalizability of all models in study 3 has declined (Figure 2) as compared to study 1 due to the geographically varying OOD test data. The training data is skewed (sample size for the 3 classes are 786, 182, & 74), which explains the poor F1 score in class 2. The zero F1 score for class 1 could be due to having only 4 test samples. Optimal models include both contextual (BERT Large FT and CT-BERT) and non-contextual (FastText DA 300).

Study 4. The decline in the generalizability of all models is more pronounced in study 4, as shown in Figure 3. All models show zero F1 scores for the September-October class 2 due to zero test



(a) May – June

(b) July – August

(c) September - October

(d) November - December

Figure 3: Study 4: Avg. test accuracy (y-axis), standard deviation (top of each bar in boxes), and avg. F1 scores (shown inside the bars). Train data (reliable categories) = samples from JanuaryApril; Test data (OOD) = 4 test sets from May-December.

cases. Also, class 2 for the November-December set has only one test sample, which lowered most of the models' F1 scores. We observe that compared to study 3, both contextual and noncontextual models perform poorly due to temporally-varying OOD test data. Their average test accuracies remain mostly below 80%. This may indicate that the temporally-varying OOD data goes through a larger shift in distribution compared to the geographically-varying OOD data.

We further observe that all models' generalizability decreases more over the future periods. For example, their November-December performance is much worse than the May-June performance. There is a substantial drop in the F1 scores of the minority classes (e.g., class 1). It indicates that may be more distribution-shift has occurred during the later months, hence models were not able to capture the language change well. The performance drop in BERT DA techniques (FT and FB) is more than BERT DS FB learning techniques. Fine-tuning on the BERT Base PTM always yields the worst generalization of all. The DS CT-BERT exhibits the best generalization performance. On the November-December data, it obtains the highest F1 score for class 1.

However, the non-contextual models do not generalize well on class 1.

Discussion.

Various techniques have been proposed to handle noisy-labeled text data that includes loss correction [42] and architectural modification [43]. However, these approaches create noisy samples by artificially corrupting samples, e.g, by uniform label flipping and random label flipping [43]. In our study, we investigate another dimension of the label-unreliability issue. Instead of randomly corrupting labels across the existing categories, we introduce a new category in the dataset that is based on samples randomly collected from the existing categories. We "pretend" that we do not have domain knowledge to determine the veracity status (misinformation or not, or what type of misinformation) of the samples from this new category. The inclusion of an "unverified" category is not an artifact in the context of COVID-19 misinformation detection problem [16]. Our intention was to see how the deep contextual models perform in presence of a noisy category. We observe more than a 20% drop in all models' performance compared to the benchmark study. We argue that we have yet to design a new class of techniques for handling noisy categories, as we did with noisy labels [42, 43].

Previously deep contextual models' robustness on OOD data was studied and it was shown that BERT-based models were more generalizable than shallow non-contextual models (e.g., Word2Vec) [14]. However, our results are contrary to this observation. We find that both the contextual and non-contextual models' generalizability vary based on the degree of distribution shift in the OOD data.

We capture the variation in distribution shift by using two orthogonal OOD datasets: variation along the spatial dimension (study 3) and variation along the temporal dimension (study 4). In addition to this, we capture the increased degree of variation within the space of the temporallyvarying data (i.e., by using 4 tests sets in study 4).

Results from studies 3 and 4 reveal some useful insights. First, all models generalize poorly in study 4 as compared to study 3. *Could this be due to the larger training set in study 3 (study 3 samples = 1042, study 4 samples = 658)?* In the technical appendix, we provide additional results (from study 5) showing that better generalization in study 3 is not due to its larger training set. *What if we could increase the training set in study 4? Would that improve the models' generalizability?* We conduct a variation of study 4 by using a larger training set (i.e.,

study 6, reported in the technical appendix). We create this training set by combining samples from January to October (1146 samples). Then, models are evaluated on the November-December OOD test set. However, we did not observe significant increase in the models' generalizability.

Results from study 3 to study 6 offer useful insights into the nature of the distribution shift in the two OOD datasets: (i) there might be a larger shift in distribution along the temporal dimension than the spatial dimension and (ii) the distribution shift becomes larger in the temporally-varying data as the temporal distance between the training and test data increases. We evaluate the contextual and non-contextual models at the backdrop of this diverse OOD space.

We see that when models were tested on the spatially-varying OOD data (study 3), the BERT-based models did not generalize better than the non-contextual models. In the case of the temporallyvarying OOD data (study 4), we find only the DS CT-BERT model to exhibit some generalizability on the November-December test set. However, on the 3 test sets from May-October, when distribution shift is comparatively smaller, non-contextual models exhibit competitive performance.

Thus, we argue that to acquire a deep understanding of a model's generalizability, we must consider the diverse nature of the OOD data.

Two pertinent questions arise on the BERT-based models' poor generalizability.

- Question 1: Why do these models exhibit the worst generalizability in study 4 as compared to study 3?
- Question 2: Why do these models' generalizability is notbetter, and sometimes worse, than the non-contextual models?

One possible answer to question 1 is that while the BERT-based models can learn contextual representations within the static space of the source data,they are not good at understanding the language change. This could explain why all BERT-based models performed worst in study 4 (data is OOD due to shifting distribution along the temporal dimension) as compared to study 3 (data is OOD due to a shift in misinformation narrative across geographic locations). Though only the CT-BERT showed good performance on the November-December data in study 4, its generalizability is significantly poor as compared to its performance in study 1 and study 3. We argue that its understanding of the shifting language space is not very deep. This could be due to the nature of the pretraining data that was limited within the initial four months (i.e., January-April of 2020) since the pandemic began. Thus, this PTM captured only as much language change that was present during that narrow time frame.

For question 2, one possible explanation is that the priors (learned from the source data) of both the DA and DS BERT models (except the CT-BERT) are much stronger than those of the noncontextual models, which may have imposed heavy inertia on its adaptation capability in the latent space for capturing the language change present in the target data.

5. CONCLUSION

In this paper, we examine the generalizability of the deep contextual models along two dimensions: (i) in presence of noisy categories and (ii) on OOD data containing varying degree of distribution shift. A systematic set of studies involving various BERT PTM-based approaches show that the deep contextual models do not generalize well when data contains unreliable categories and is OOD. We explain the lack of generalizability of the deep contextual models.

In the future, we will include additional contextual models in our generalizability investigation that are (i) Transformer-based non-BERT and (ii) not Transformer based.

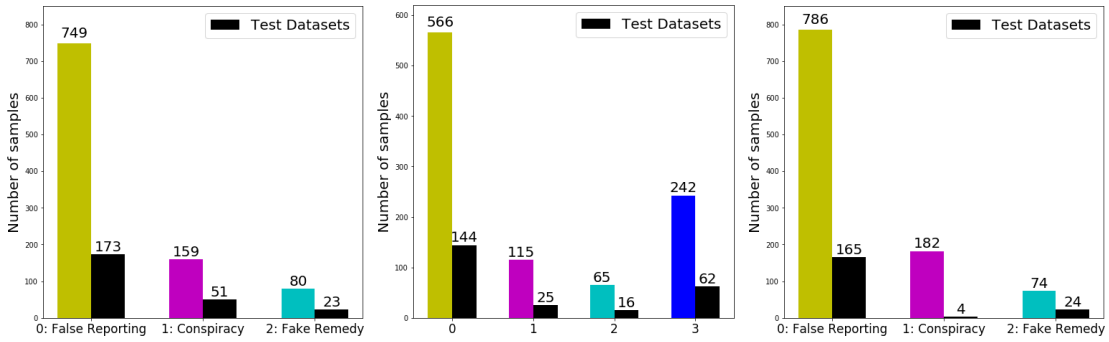
APPENDIX

In this Appendix, we present the sample distribution per class for all studies. Then, we describe the design of two data ablation studies and analyze the results.

Sample Distribution in the Studies

In this section, we present the sample distribution for both the training and test samples for all studies.

Study 1, 2, & 3 Figure 4 shows that in studies 1, 2, and 3, the sample distribution is heavily skewed, i.e., the number of samples in the "False Reporting" class is significantly larger than other classes. In study 2, the "Unverified" category (class 3) contains more samples than classes 1 and 2.



(a) Study 1 (b) Study 2 (c) Study 3

Figure 4: Sample distribution.

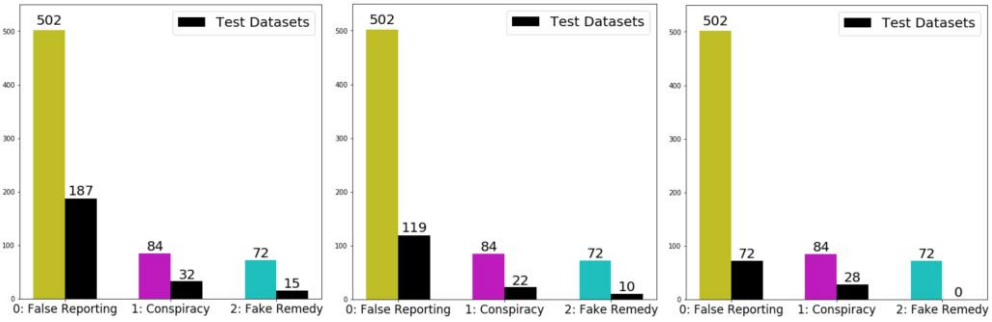
In study 3 (Figure 4(b)), the number of test samples from class 2 is only 4, which explains why all models obtain an average zero F1 score on class 2.

Study 4 In study 4, the training dataset is created using samples from January-April of 2020, while the four OOD test sets are created using samples from May-June, July-August, SeptemberOctober, and November-December, respectively. Sample distribution for both the training and all test datasets are shown in Figure 5. Similar to the previous three studies, the datasets are heavily skewed. The test samples in some classes in the September-October and November-December datasets are scarce. For example, class 2 in the September-October dataset has no samples and class 3 in the November-December dataset has only one sample. This explains why we could not reliably evaluate the performance of the models on class 2 (September-October dataset) and class 3 (November-December dataset).

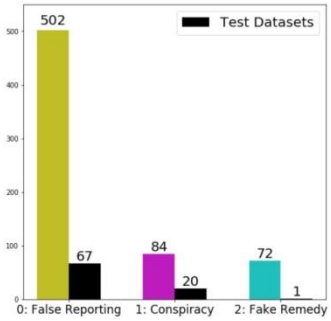
Design of the Data Ablation Studies

We design two data "ablation" studies by editing the existing datasets. The goal is to corroborate the observed performance of the models on the OOD data in studies 3 and 4. In study 3, models were evaluated on the geographically varying OOD data, and in study 4, the evaluation was based on the temporally varying OOD data for various degrees of the distribution shift. We

observed that models were more generalizable on the geographically varying OOD data (study 3) than in the case of the temporally varying OOD data. To determine whether better generalizability in

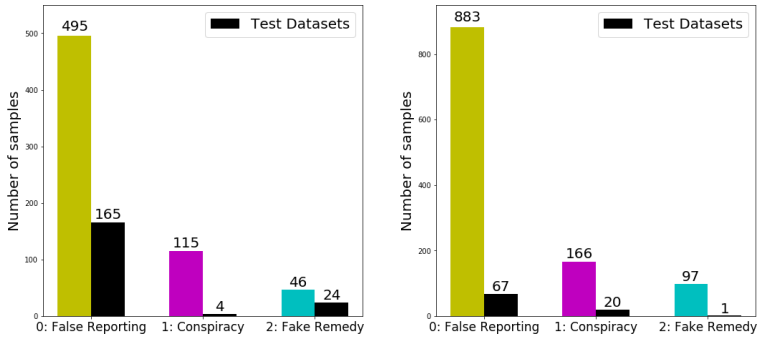


(a) May-June (b) July-August (c) September-October



(d) November-December

Figure 5: Sample distribution for the four datasets in study 4.



(a) Study 5 (b) Study 6

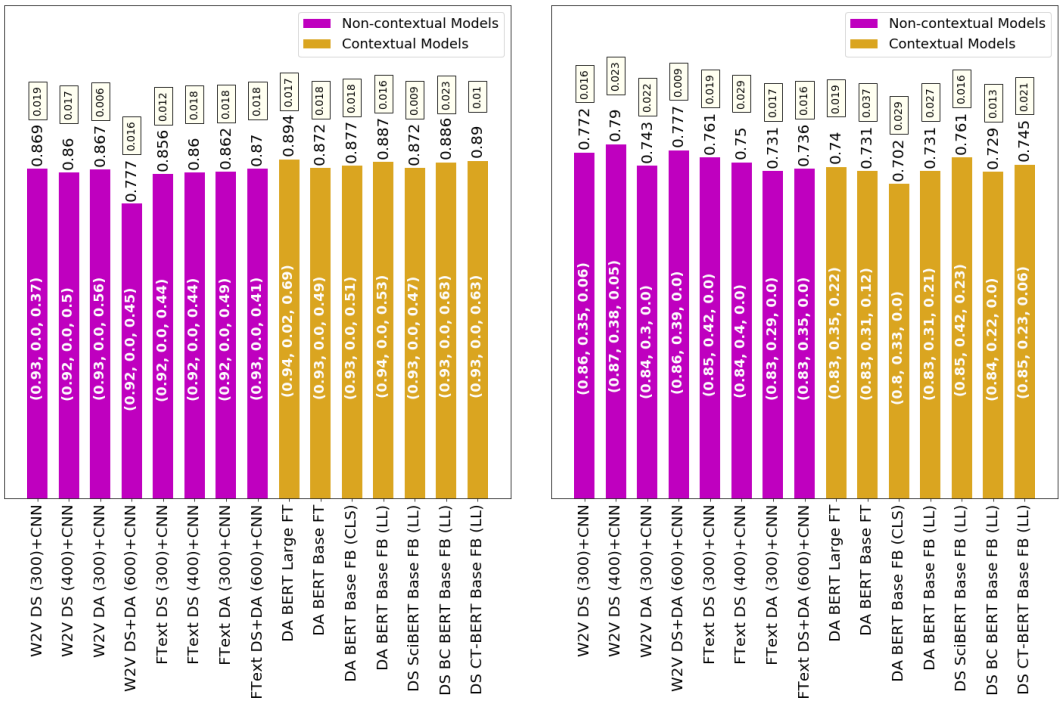
Figure 6: Sample distribution.

study 3 is due to the larger size of the training data, we design a new study (study 5).

In addition to this, we observed that in study 4, the models generalized extremely poorly on the September-December test datasets compared to the test datasets from the earlier months MayAugust. To determine whether the poor performance during later months is due to the smaller training set (i.e., the January-April training set) or due to the larger shift in the test data, we design a study (study 6).

These two new studies are intended to explain away the effect of the size of the training set in studies 3 and 4.

Study 5: Explaining Away the Effect of Larger Training Set in Study 3 The size of the training data in study 2 (i.e., 1042) is much larger than the training set in study 3 (i.e., 658). However, we



(a) Study 5. (b) Study 6.

Figure 7: Avg. test accuracy (y-axis), standard deviation (top of each bar in boxes), and avg. F1 scores (shown inside the bars). “CLS”: embeddings of the classification token; “LL”: embeddings from the last hidden layer.

believe that better generalization in study 3 is not due to its larger training set. Thus, for explaining away the effect of a larger training set in study 3, we reduce its size so that it becomes equal to the size of the training set in study 4. The sample distribution for study 5 is shown in Figure 6(a).

Study 6: Explaining Away the Effect of Smaller Training Set in Study 4 We find that models generalize poorly on all four test sets in study 4. The performance degradation is severe in the September-October and November-December test sets. We believe that this decline in generalization is not due to the smaller training set, but because of the nature of the OOD test data, i.e., data from the distant future (e.g., samples from September-December) exhibit more distribution shift compared to the near future (e.g., samples from May-August)

For explaining away the effect of the size of the training set in study 4, we increase its size by combining samples from January to October. Then, models trained using this large set of 1146 samples are evaluated on the November-December OOD test set. The sample distribution for study 6 is shown in Figure 6(b).

Results of the Data Ablation Studies Below we present the results obtained from the two data ablation studies.

Study 5. Figure 7(a) shows the results from study 5. We observe that although the size of the training set is reduced from 1042 to 656, the performance of the models either remained unchanged or increased, which explains away the effect of the size of the training set in study 3.
Study 6. Figure 7(b) shows the results obtained from the study 6. Even after increasing the size of the training set, the performance gain of the best contextual model from study 4, i.e., the CTBERT Base FB (last layer), is minor (its test accuracy increased from 0.709 to 0.745). BERT FT models are benefited most from the increased data. However, none of the models exceed test accuracy above 80% and achieve above 50% average F1 score on the minority classes. Thus, we see that increasing training data by including samples from "near-future" months did not improve the generalizability of the models on the November-December OOD data, which explains away the effect of the size of the training set in study 4.

Results from studies 5 and 6 indicate that (i) temporally varying OOD data shows more distribution shift than geographically varying OOD data and (ii) the degree of the distribution shift increases as the temporal distance between the training data and the test data increases. That is why the models generalize poorly on the temporally varying OOD data in study 4 and perform significantly poorly on the November-December test data from study 4.

REFERENCES

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA (I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, eds.), pp. 5998–6008, 2017.

[2] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," J. Mach. Learn. Res., vol. 21, pp. 140:1–140:67, 2020.

[3] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," in Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, (Brussels, Belgium), pp. 353–355, Association for Computational Linguistics, Nov. 2018.

[4] R. T. McCoy, J. Min, and T. Linzen, "BERTs of a feather do not generalize together: Large variability in generalization across models with similar test set performance," in Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, (Online), pp. 217–227, Association for Computational Linguistics, Nov. 2020.

[5] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning (still) requires rethinking generalization," Commun. ACM, vol. 64, pp. 107–115, Feb. 2021.

[6] S. Fort, J. Ren, and B. Lakshminarayanan, "Exploring the limits of out-of-distribution detection," ArXiv, vol. abs/2106.03004, 2021.

[7] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang, "Pre-trained models for natural language processing: A survey," 2020.

[8] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), (New Orleans, Louisiana), pp. 2227–2237, Association for Computational Linguistics, June 2018.

[9] G. Paolacci, J. Chandler, and P. G. Ipeirotis, "Running experiments on amazon mechanical turk," Judgment and Decision Making, vol. 5, no. 5, pp. 411–419, 2010.

[10] B. Frenay and M. Verleysen, "Classification in the presence of label noise: A survey,"′ IEEE Transactions on Neural Networks and Learning Systems, vol. 25, pp. 845–869, 2014.

[11] J. Eisenstein, “What to do about bad language on the internet,” in Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (Atlanta, Georgia), pp. 359–369, Association for Computational Linguistics, June 2013.

[12] J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, Dataset Shift in Machine Learning. The MIT Press, 2009.

[13] R. Goel, S. Soni, N. Goyal, J. Paparrizos, H. M. Wallach, F. Diaz, and J. Eisenstein, “The social dynamics of language change in online networks,” in Social Informatics - 8th International Conference, SocInfo 2016, Bellevue, WA, USA, November 11-14, 2016, Proceedings, Part I (E. S. Spiro and Y. Ahn, eds.), vol. 10046 of Lecture Notes in Computer Science, pp. 41–57, 2016.

[14] D. Hendrycks, X. Liu, E. Wallace, A. Dziedzic, R. Krishnan, and D. Song, “Pretrained transformers improve out-of-distribution robustness,” in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020 (D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, eds.), pp. 2744–2751, Association for Computational Linguistics, 2020.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “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), (Minneapolis, Minnesota), pp. 4171–4186, Association for Computational Linguistics, June 2019.

[16] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, “Deep learning based text classification: A comprehensive review,” 2021.

[17] S. A. Memon and K. M. Carley, “Characterizing COVID-19 misinformation communities using a novel twitter dataset,” CoRR, vol. abs/2008.00791, 2020.

[18] M. Cheng, S. Wang, X. Yan, T. Yang, W. Wang, Z. Huang, X. Xiao, S. Nazarian, and P. Bogdan, “A covid-19 rumor dataset,” Frontiers in Psychology, vol. 12, p. 1566, 2021.

[19] S. Siwakoti, K. Yadav, I. Thange, N. Bariletto, L. Z. andAlaaGhoneim, and J. N. Shapiro, “Localized misinformation in a global pandemic: Report on covid-19 narratives around the world,” Empirical Study of Conflict, pp. 1–68, 2021.

[20] N. F. Johnson, N. Velasquez, O. K. Jha, H. Niyazi, R. Leahy, N. J. Restrepo, R. Sear, P. Manrique, Y. Lupu, P. Devkota, and S. Wuchty, “Covid-19 infodemic reveals new tipping point epidemiology and a revised r formula,” 2020.

[21] L. Cui and D. Lee, “Coaid: COVID-19 healthcare misinformation dataset,” CoRR, vol. abs/2006.00885, 2020.

[22] T. Hossain, R. L. Logan IV, A. Ugarte, Y. Matsubara, S. Young, and S. Singh, “COVIDLies: Detecting COVID-19 misinformation on social media,” in Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020, (Online), Association for Computational Linguistics, Dec. 2020.

[23] H. Y. Lin and T.-S. Moh, Sentiment Analysis on COVID Tweets Using COVID-Twitter-BERT with Auxiliary Sentence Approach, pp. 234–238. New York, NY, USA: Association for Computing Machinery, 2021.

[24] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” CoRR, vol. abs/1301.3781, 2013.

[25] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching word vectors with subword information,” arXiv preprint arXiv:1607.04606, 2016.

[26] I. Beltagy, K. Lo, and A. Cohan, “SciBERT: A pretrained language model for scientific text,” in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLPIJCNLP), (Hong Kong, China), pp. 3615–3620, Association for Computational Linguistics, Nov. 2019.

[27] M. E. Peters, S. Ruder, and N. A. Smith, “To tune or not to tune? adapting pretrained representations to diverse tasks,” in Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019), (Florence, Italy), pp. 7–14, Association for Computational Linguistics, Aug. 2019.

[28] Y. Kim, “Convolutional neural networks for sentence classification,” in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), (Doha, Qatar), pp. 1746–1751, Association for Computational Linguistics, oct 2014.

[29] M. Muller, M. Salath¨ e, and P. E. Kummervold, “Covid-twitter-bert: A natural language pro-´ cessing model to analyse COVID-19 content on twitter,” CoRR, vol. abs/2005.07503, 2020.
[30] A. Tendle and M. R. Hasan, “A study of the generalizability of self-supervised representations,” Machine Learning with Applications, vol. 6, p. 100124, 2021.
[31] S. Brennen, F. N. Simon, P. K. Howard, and R. u. Nielsen, “Types, sources, and claims of covid-19 misinformation,” Reuters Institute for the Study of Journalism, Apr 2020.
[32] S. J. Pan and Q. Yang, “A survey on transfer learning,” IEEE Transactions on knowledge and data engineering, vol. 22, no. 10, pp. 1345–1359, 2009.
[33] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, “Transformers: State-of-the-art natural language processing,” in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, (Online), pp. 38–45, Association for Computational Linguistics, Oct. 2020.
[34] C. May, A. Wang, S. Bordia, S. R. Bowman, and R. Rudinger, “On measuring social biases in sentence encoders,” 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), (Minneapolis, Minnesota), pp. 622–628, Association for Computational Linguistics, June 2019.
[35] Y. Qiao, C. Xiong, Z. Liu, and Z. Liu, “Understanding the behaviors of BERT in ranking,” CoRR, vol. abs/1904.07531, 2019.
[36] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence embeddings using Siamese BERTnetworks,” in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), (Hong Kong, China), pp. 3982–3992, Association for Computational Linguistics, Nov. 2019.
[37] E. Alsentzer, J. R. Murphy, W. Boag, W. Weng, D. Jin, T. Naumann, and M. B. A. McDermott, “Publicly available clinical BERT embeddings,” CoRR, vol. abs/1904.03323, 2019.
[38] Google, “Google code archive: word2vec.” https://code.google.com/archive/ p/word2vec/, 2013. Accessed: 2025-9-12.
[39] FastText, “Fasttext.” https://fasttext.cc/docs/en/english-vectors. html, 2020. Accessed: 2025-9-12.
[40] R. Reh˘ u˚˘rek and P. Sojka, “Software Framework for Topic Modelling with Large Corpora,” in Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, (Valletta, Malta), pp. 45–50, ELRA, May 2010.
[41] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, “On the variance of the adaptive learning rate and beyond,” ArXiv, vol. abs/1908.03265, 2020.
[42] G. Patrini, A. Rozza, A. K. Menon, R. Nock, and L. Qu, “Making deep neural networks robust to label noise: A loss correction approach,” in 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, pp. 2233– 2241, IEEE Computer Society, 2017.
[43] I. Jindal, D. Pressel, B. Lester, and M. S. Nokleby, “An effective label noise model for DNN text classification,” in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACLHLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers) (J. Burstein, C. Doran, and T. Solorio, eds.), pp. 3246–3256, Association for Computational Linguistics, 2019.

AUTHOR

Dr. M. R. Hasan is an Assistant Professor in the Department of Electrical and Computer Engineering at the University of Nebraska-Lincoln, specializing in Artificial Intelligence (AI) and Big Data. Originally trained as a theoretical physicist, he earned his Ph.D. in Computing and Information Systems from the University of North Carolina at Charlotte and now directs the Human-First Artificial Intelligence Lab (HAL 2.0). His research focuses on developing trustworthy multimodal AI systems that integrate diverse sources of information to advance applications in education, healthcare, climate resilience, and algorithmic trust. Dr. Hasan has been recognized with the 2025 ACM International Conference on Multimodal Interaction (ICMI) Blue Sky Award for pioneering contributions to AI and human experience modeling and has led NSF-funded projects advancing socially impactful AI technologies. He also contributes to institutional leadership as a member of the University of Nebraska System AI Task Force and organizer of initiatives supporting AI-driven transformation across campus.