Overcoming Language Disparity in Online Content Classification with Multimodal Learning

Gaurav Verma, Rohit Mujumdar, Zijie J. Wang, Munmun De Choudhury, Srijan Kumar

Georgia Institute of Technology
{gverma, rohitmujumdar, jayw, munmund, srijan}@gatech.edu

Abstract

Advances in Natural Language Processing (NLP) have revolutionized the way researchers and practitioners address crucial societal problems. Large language models are now the standard to develop state-of-the-art solutions for text detection and classification tasks. However, the development of advanced computational techniques and resources is disproportionately focused on the English language, sidelining a majority of the languages spoken globally. While existing research has developed better multilingual and monolingual language models to bridge this language disparity between English and non-English languages, we explore the promise of incorporating the information contained in images via multimodal machine learning. Our comparative analyses on three detection tasks focusing on crisis information, fake news, and emotion recognition, as well as five high-resource non-English languages, demonstrate that: *(a)* detection frameworks based on pre-trained large language models like BERT and multilingual-BERT systematically perform better on the English language compared against non-English languages, and *(b)* including images via multimodal learning bridges this performance gap. We situate our findings with respect to existing work on the pitfalls of large language models, and discuss their theoretical and practical implications.

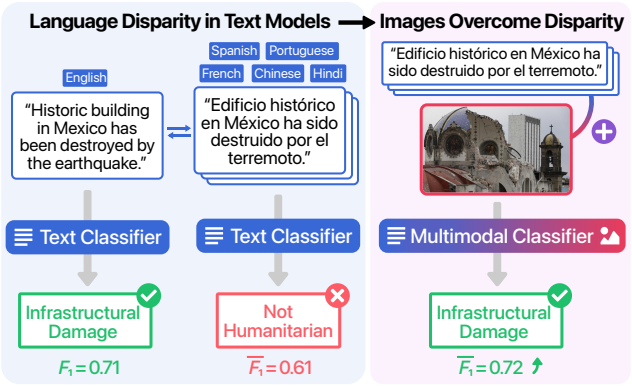


Figure 1: Overview figure. We use multimodal (image + text) learning to overcome the language disparity that exists between English and non-English languages. The figure illustrates an example of a social media post that is correctly classified in English but misclassified in Spanish. Including the corresponding image leads to correct classification in Spanish as well as other non-English languages.

Introduction

Users of social computing platforms use different languages to express themselves (Mocanu et al. 2013). These expressions often give us a peek into personal-level and societal-level discourses, ideologies, emotions, and events (Kern et al. 2016). It is crucial to model *all* of these different languages to design equitable social computing systems and to develop insights that are applicable to a wider segment of the global population.

In recent years, we have seen remarkable ability in using linguistic signals and linguistic constructs extracted from social media and web activity toward tackling societal challenges, whether in detecting crisis-related information (Houston et al. 2015) or identifying depression-related symptoms (De Choudhury et al. 2013). While earlier approaches relied on qualitative language inference techniques (Crook et al. 2016), using pre-existing dictionaries (Pennebaker, Francis, and Booth 2001), and traditional classifiers (Glasgow, Fink, and Boyd-Graber 2014), more recent approaches leverage the advances in deep learning-based language modeling techniques. Large pre-trained models like BERT (Devlin et al. 2018) are frequently used to train classifiers in tasks pertaining to social good (Singhal et al. 2019; Sun, Huang, and Qiu 2019) and are now a new standard to build state-of-the-art classification systems to support real-world decision-making.

As Joshi et al. (2020) illustrate, these rapidly evolving language technologies and their applications are largely focused on only a very small number of over 7000 languages in the world. A majority of the research in natural language processing (NLP) is focused on a few high resource languages, and disproportionately on English (Mielke 2016; Bender 2019). The development of systems that can model languages beyond English is important for ensuring *(a)* inclusion of communities, *(b)* equitable extension of services that are driven by these language technologies to diverse groups, and *(c)* preservation of endangered languages (Muller et al. 2021). Especially in the context of social computing, language-specific lapses can lead to in-

equitable outcomes. For instance, lower detection abilities on Twitter posts published in Spanish could possibly lead to inequitable humanitarian interventions in times of crisis; and, the lack of powerful misinformation detectors for the Chinese language can possibly lead to situations where specific-language speaking individuals are more vulnerable to health-related misinformation. As BERT-like monolingual and multilingual models take a central role in building approaches to address crucial societal tasks, the bias toward the English language can propagate, reinforce, and even exacerbate the existing inequities that many underserved groups face (PewResearch 2018).

Existing attempts to bridge this gap between English and non-English languages have focused on developing better multilingual and monolingual (non-English) language models (Nozza, Bianchi, and Hovy 2020). In this work, we explore the promise of information that lies in other complementary modalities, specifically images (1). Considering images as an additional modality has proven to be beneficial in a wide range of scenarios — from accurately estimating dietary intake in a pediatric population (Higgins et al. 2009), to creating effective questionnaires (Reynolds and Johnson 2011). The underlying idea stems from the simple fact that images are not bound by any language. We propose the use of multimodal learning, which jointly leverages the information in related images and text, to boost performance on the non-English text and effectively bring it closer to the performance on English text. More concretely, we study the following two research questions in this work:

RQ1: *Does using large language models for social computing tasks lead to lower performance on non-English languages when compared to the English language?*

RQ2: *Can inclusion of images with multimodal learning help in bridging the performance gap between English and non-English models?*

To this end, we study the performance of fine-tuned BERT-based monolingual models and multilingual-BERT on three distinct *classification tasks* that are relevant to social computing: *(i)* humanitarian information detection during crisis (Ofli, Alam, and Imran 2020), *(ii)* fake news detection (Shu et al. 2017), and *(iii)* emotion detection (Duong, Lebret, and Aberer 2017). These tasks involve categorizing posts/articles published on the web into real-world concepts that help determine, for instance, the type of humanitarian effort required during a crisis or the veracity of published news. Besides English, we consider five high-resource languages: Spanish, French, Portuguese, (Simplified) Chinese, and Hindi. Via extensive comparative analysis on these existing datasets, we demonstrate that *(a)* large language models — whether monolingual or multilingual — systematically perform better on English text compared to other high-resource languages, and *(b)* incorporating images as an additional modality leads to considerably lesser deviation of performance on non-English languages with respect to that on English[1]. We conclude by discussing the implications of these findings from both practical and theoretical standpoints, and situate them with respect to prior knowledge from the domains of NLP and social computing.

[1]Project webpage with resources: https://multimodality-language-disparity.github.io/

Related Work

We discuss three major themes of research that are relevant to our work: the use of large language models in developing approaches for social computing tasks, the discussion of the pitfalls of large language models and their treatment of non-English languages, and the role of multimodal learning in developing social media classification systems.

Large language models for social computing tasks: Development and deployment of large language models — deep learning models trained on massive amounts of data collected from the web, have transformed not only the field of NLP but also related fields that leverage text data to make inferences (Rasmy et al. 2021). To this end, large language models have been used for various applications in social computing (Arviv, Hanouna, and Tsur 2021; Choi et al. 2021). The effectiveness of language models in addressing these tasks can be primarily attributed to two factors: (i) they are trained on massive amounts of unannotated text data, leading to a general understanding of natural language, and (ii) they can be easily fine-tuned for specific tasks with moderately-sized annotated data to demonstrate task-specific understanding. Several language models such as BERT (Devlin et al. 2018) and T5 (Raffel et al. 2020) have been developed for the English language. Since these models cover only English, large multilingual variants like mBERT (Devlin et al. 2018) and mT5 (Xue et al. 2021) have also been developed to model over a hundred other languages beyond English. These language models (both monolingual and multilingual) are widely adopted to develop state-of-the-art approaches for several tasks where the textual modality withholds key information.

Language disparity in NLP: Scholars have discussed the disproportionate focus in NLP research on the English language (Bender 2019; Joshi et al. 2020; Mielke 2016). Since approaches to address social computing tasks are increasingly relying on NLP techniques centered around large language models, it is important to understand the possible implications of this disproportionate focus on the state of social computing research. Prior studies have tried to understand the pitfalls of using large language models — environmental and financial costs (Strubell, Ganesh, and McCallum 2019), reliance on data that represents hegemonic viewpoints (Bender et al. 2021), encoding biases against marginalized populations (Basta, Costa-jussà, and Casas 2019). However, our work focuses on comparing English language models with non-English language models in a social computing context. Similar to English, multilingual variants of language models are used to develop the state-of-the-art[2] approaches for multiple high-resource non-English languages (Nozza, Bianchi, and Hovy 2020). To this end, previous research has focused on understanding how multilingual language models treat various non-English languages relative to each other, especially the contrast between high-resource and low-resource

[2]Leaderboard: https://bertlang.unibocconi.it/

Figure 2: Illustrative examples from considered multimodal datasets. We consider three classification datasets for our experiments: (A) crisis humanitarianism dataset (number of classes: 5; infrastructure and utility damage: 10%, rescue volunteering or donation effort:14%, affected individuals: 1%, other relevant information:22%, & not humanitarian:53%), (B) fake news detection dataset (number of classes: 2; fake:21% & real:79%), and (C) emotion detection dataset (number of classes: 4; creepy:22%, rage:19%, gore:25%, & happy:34%).

languages (Pires, Schlinger, and Garrette 2019; Wu and Dredze 2020; Nozza, Bianchi, and Hovy 2020; Muller et al. 2021). In this work, we do not focus on the general pitfalls of large language models or comparisons across non-English languages. Instead, we aim to establish the language disparity between English and non-English languages that is caused due to the adoption of large language models.

Multimodal learning: Multimodal learning involves relating information from multiple content sources. On the web, the text is often associated with images, especially on social media platforms like Twitter, Instagram, and Facebook. Multimodal learning allows us to combine modality-specific information into a joint representation that captures the real-world concept corresponding to the data (Ngiam et al. 2011). To this end, inference based on multimodal learning has demonstrated better performance than both text-only and image-only methods, especially in scenarios where access to complementary information can be crucial (e.g., assessing whether a Twitter post (image + text) is about disaster (Ofli, Alam, and Imran 2020), or if a news article (image + title) is fake (Singhal et al. 2020), whether the Reddit post conveys rage (Duong, Lebret, and Aberer 2017)). However, the studies that demonstrate the effectiveness of multimodal learning do so while making comparisons against language-specific text-only methods, without making any comparisons across different languages. In this work, we aim to use multimodal learning, more specifically images, to bridge the gap between English and non-English languages.

Datasets

To achieve robust and generalizable findings, we utilize a comparative analytic approach on three different pre-existing datasets that cover issues like humanitarian information processing, fake news detection, and emotion detection. Figure 2 presents some examples from the three datasets discussed below as well as the proportion of classes.

Multimodal crisis humanitarian dataset: In times of crises, social media often serves as a channel of communication between affected parties and humanitarian organizations that process this information to respond in a timely and effective manner. To aid the development of computational methods that can allow automated processing of such information and, in turn, help humanitarian organizations in gaining real-time situational awareness and planning relief operations, Alam et al. (Alam, Ofli, and Imran 2018) curated the CrisisMMD dataset. This multimodal dataset comprises $7,216$ Twitter posts (images + text) that are categorized into 5 humanitarian categories. The dataset covers 7 crises that occurred in 2017 all over the globe (3 hurricanes, 2 earthquakes, 1 wildfire and floods). We formulate the task of humanitarian information detection as a multi-class classification problem, and use the standardized training ($n = 5263$), evaluation ($n = 998$), and test ($n = 955$) sets in our experiments. We maintain the exact same training, validation, and test splits for all the experiments that involve this dataset.

Multimodal fake news dataset: Ease of publishing news on online platforms, without fact-checking and editorial rigor, has often led to the widespread circulation of misleading information on the web (Lazer et al. 2018). Shu et al. (2017; 2018) curated the FakeNewsNet dataset to promote research on multimodal fake news detection; it comprises full-length news articles (title + body) from two different domains: politics (fake/real labels provided by PolitiFact) and entertainment (fake/real labels provided by GossipCop) and the corresponding images in the articles. The fake news detection task can therefore be formulated as a binary classification task, where the `label:0` corresponds to the `real` class and the `label:1` corresponds to the `fake` class. We use the pre-processed version of the dataset provided by Singhal et al. (2020) and consider only the title of the news article for our experiments while dropping the body of the article. Furthermore, we combine the two domains (entertainment and politics) to create a single dataset and use the same train and test splits like Singhal et al. We, however, randomly split the original train set in $90 : 10$ ratio to create an updated train and validation set. Effectively, our final train, validation, and

Language	Fluency	Meaning	Cohen's κ
Spanish (es)	4.01	4.10	0.81
French (fr)	4.07	4.24	0.83
Portuguese (pt)	3.98	4.22	0.86
Chinese (zh)	4.06	4.29	0.84
Hindi (hi)	3.91	4.12	0.82

Table 1: Quality assessment of *machine* translation. Average scores assigned by human annotators on a 5-point Likert scale ($1-5$) for translation quality of generated text, and the agreement scores between annotators for each language for the fluency scores. $N = 200$ examples per language per dataset; 3 annotators per example.

Language	Fluency	Meaning	Cohen's κ
Spanish (es)	4.21	4.33	0.85
French (fr)	4.19	4.29	0.82
Portuguese (pt)	4.08	4.36	0.79
Chinese (zh)	4.31	4.40	0.85
Hindi (hi)	4.39	4.45	0.87

Table 2: Quality assessment of *human* translation for crisis humanitarianism dataset. Average scores assigned by human annotators on a 5-point Likert scale ($1-5$) for translation quality of generated text, and the agreement scores between annotators for each language for the fluency scores. $N = 200$ examples per language; 3 annotators/example.

test sets comprise 9502, 1055, and 2687 news articles, each example containing the title of the news and an image.

Multimodal emotion dataset: Using user-generated content on the web to infer the emotions of individuals is an important problem, with applications ranging from targeted advertising (Teixeira, Wedel, and Pieters 2012) to detecting mental health indicators (De Choudhury, Counts, and Horvitz 2013). To this end, we collect the dataset introduced by Duong, Lebret, and Aberer (2017) for the task of multimodal emotion detection. The dataset comprises Reddit posts categorized into 4 emotion-related classes, `creepy`, `gore`, `happy`, and `rage`, where each post contains an image and text. We crawled the images from Reddit using the URLs provided by the authors and randomly split the dataset in a 80:10:10 ratio to obtain the train ($n = 2568$), validation ($n = 321$), and test ($n = 318$) sets. Similar to other datasets, we maintain the exact same splits for all the experiments that involve this dataset to ensure consistent comparisons.

Curating non-English datasets: All the three datasets discussed above only have texts (Twitter posts, news articles, and Reddit posts) in English. Given the lack of non-English multimodal datasets, we employ machine translation to convert English text into different target languages. For translation, we use the MarianNMT system, which is an industrial-grade machine translation system that powers Microsoft Translator (Junczys-Dowmunt et al. 2018). As target languages, we consider the following five non-English languages: Spanish (`es`), French (`fr`), Portuguese (`pt`), Simplified Chinese (`zh`), and Hindi (`hi`). Together, these five languages represent culturally diverse populations – minority groups in the United States (Hispanics), Asians, and the Global South, and are written in various scripts – Latin (`es`, `fr`, and `pt`), Hanzi (`zh`), and Devanagari (`hi`). It is worth noting that none of these five non-English languages are considered to be low-resource languages (Hedderich et al. 2021) – which is a more appropriate designation for languages like Sinhala, the Fijian language, and Swahili. However, since these languages are sufficiently high-resource languages, MarianNMT can produce high-quality translations in these languages from the original English text.

We use the pre-trained language-specific translation models of MarianNMT, made available via HuggingFace (Wolf et al. 2019), to translate the text part of each example in the three datasets to the five target language (`en` $\rightarrow$ `es`, `fr`, `pt`, `zh`, `hi`). Before translating, we pre-processed the English text to remove URLs, emoticons, platform-specific tokens (like 'RT' for indicating retweets on Twitter), and symbols like @ and #. We also expanded negatives like *can't* and *won't* to '*can not*' and '*will not*'. Overall, translating the English text to five non-English languages gives us 6 different versions of each of the three datasets discussed above, where each version differs only in terms of the language of the text. It is worth emphasizing that the train, validation, and test splits remain the same across different languages; this is done to ensure a meaningful comparison of classification models' performance across different languages.

Human-translated subset for crisis humanitarianism: Besides the machine-translated text, we also obtain manual translations for a subset of examples from the test set of the Crisis Humanitarianism dataset. For Spanish, French, and Portuguese, we recruited workers from Amazon Mechanical Turk (AMT) who were designated as 'Masters' and proficient in both English and the target language. For Chinese and Hindi, we obtained annotations from doctoral students fluent in both English and Chinese/Hindi. The recruited participants translated 200 examples from the test set for each non-English language. The annotators were shown both the original Twitter post and were instructed to translate the text to the target language while maintaining grammatical coherence and preserving semantic meaning. We use this manually-translated subset of the test set for evaluation purposes alone — allowing us to observe the validity of observed trends on a cleaner dataset. Next, we assess the quality of machine- and human-translated text.

Human evaluation of translation quality: MarianNMT is the engine behind Microsoft Translator, a system that demonstrates translation quality that is close to human parity for specific languages and in constrained settings (Microsoft 2019). We conduct an independent evaluation of the generated translations of examples from our datasets. For this, we randomly sampled 200 examples from each dataset (600 examples in total) and asked human annotators to assess the translation quality. Similar to above, the recruited annotators were AMT workers for Spanish, French, and Portuguese, and doctoral students for Chinese and Hindi. Each of the 3000 (i.e., 600×5) translation pairs was annotated by 3 annotators where they responded to the following two questions using a five-point Likert scale (1: strongly dis-

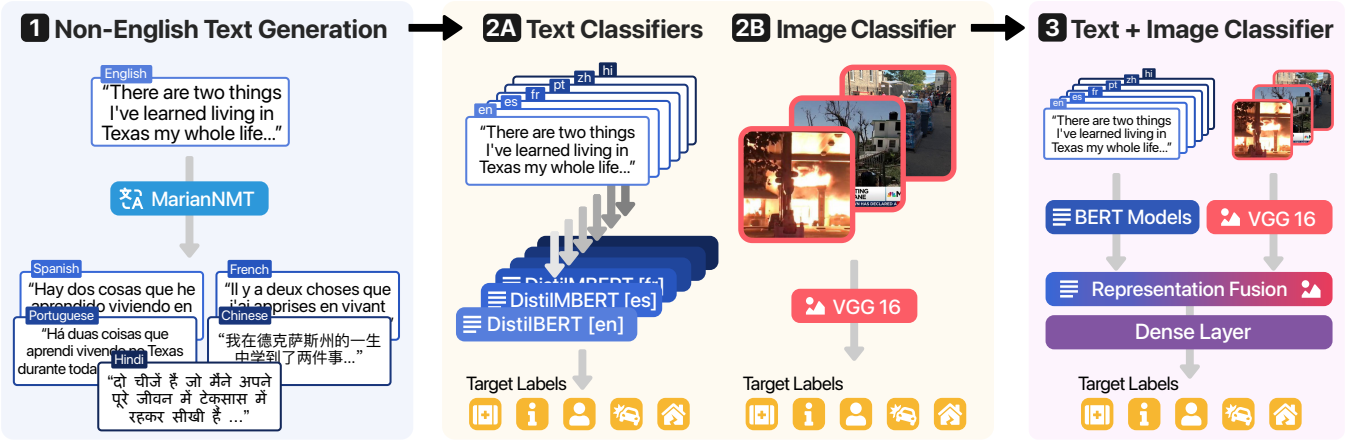


Figure 3: Overview of the adopted methodology. After using machine translation to obtain high-quality translations of the English text in our datasets (1), we train language-specific text-only classification models (2A) and image-only classification models (2B). The multimodal classifier (3) fuses the representations obtained from trained text-only and image-only models, and predicts the label based on joint modeling of the input modalities.

agree, ..., 5: strongly agree): (i) *Is the <Spanish>*[3] *text a good translation of the English text?*, and (ii) *Does the <Spanish> text convey the same meaning as the English text?* While the first question encouraged the annotators (i.e., AMT workers for Spanish, French, and Portuguese, and doctoral students for Chinese and Hindi) to evaluate the quality of the translations, including grammatical coherence, the second question encouraged them to assess the preservation of meaning in the generated translation, a relatively relaxed assessment. As shown in Table 1, the annotators' responses to the first question indicate that the translation qualities were reliable. We observe high average scores on the Likert scale as well as strong inter-annotator agreements (computed using Cohen's κ) for all five languages. For the second question, the average scores on the Likert scale are consistently $\geq$ 4.10 for all the five languages, indicating almost perfect preservation of meaning after translation from the English text to the target language.

Finally, we conducted a similar assessment of the quality of the human-translated subset of the Crisis Humanitarianism dataset. Each of 1000 (i.e., 200 $\times$ 5) translation pairs were similarity annotated by 3 annotators. As expected, Table 2 shows that the fluency and meaning preservation in the human-translated text is better than the machine-translated text with strong inter-annotator agreement scores.

In the upcoming sections, we describe the training and evaluation of the classification models, and the results for RQ1 and RQ2. Figure 3 provides an overview of our method.

[3]The language name was changed as per the target language for which the annotators were rating. Also, we inserted some "attention-check" examples during the annotation process to ensure the annotators read the text carefully before responding. This was done by explicitly asking the annotators to mark a randomly-chosen score on the Likert scale regardless of the original and translated text. We discard the annotations from annotators who did not respond to all the attention-check examples correctly.

Language Disparity with Language Models

In this section, we focus on RQ1: whether using large language models for classification tasks results in a systematic disparity between the performance on English and non-English text. We use pre-trained language models and fine-tune them on the specific classification task using language-specific labeled datasets.

Classification models for English: We use two pre-trained language models: DistilBERT (Sanh et al. 2019) (`distilbert-base-cased` on HuggingFace) and DistilmBERT (`distilbert-base-multilingual-cased` on HuggingFace) to classify the English text. We fine-tune the pre-trained language models on the 3 datasets discussed above by using the respective training sets. The process of fine-tuning a language model involves taking a pre-trained language model[4] and replacing the "pre-training head" of the model with a randomly initialized "classification head". The randomly initialized parameters in the classification head are learned by *fine-tuning* the model on classification examples while minimizing the cross-entropy loss. To train the English language classification models for each dataset, we use Adam optimizer (Kingma and Ba 2014) with a learning rate initialized at 10^{-4}; hyper-parameters are set by observing the classification performance achieved on the respective validation set. We use early stopping (Caruana, Lawrence, and Giles 2000) to stop training when the loss value on the validation set stops to improve for 5 consecutive epochs.

Classification models for non-English languages: To classify the non-English languages into task-specific categories, we use two set of pre-trained language models: *(a)* monolingual models and *(b)* multilingual model called DistilmBERT (`distilbert-base-multilingual-cased` on Hug-

[4]Large language models are typically pre-trained in a self-supervised manner (e.g., predicting a masked word, given other surrounding words that contextualize the masked word (Devlin et al. 2018)) using corpora that comprises billions of words.

Language - Model	Crisis Humanitarianism				Fake News Detection				Emotion Detection			
	F_1	Precision	Recall	Accuracy	F_1	Precision	Recall	Accuracy	F_1	Precision	Recall	Accuracy
Monolingual BERTs												
English - DisilBERT	0.71	0.72	0.70	0.80	0.59	0.64	0.56	0.85	0.79	0.79	0.79	0.80
Spanish - BETO	0.64	0.67	0.63	0.78	0.54	0.63	0.47	0.84	0.75	0.76	0.75	0.77
French - CamemBERT	0.69	0.69	0.69	0.77	0.56	0.60	0.53	0.84	0.76	0.76	0.76	0.78
Portuguese - BERTimbau	0.67	0.67	0.68	0.77	0.57	0.58	0.55	0.84	0.71	0.72	0.70	0.73
Chinese - ChineseBERT	0.65	0.64	0.66	0.75	0.56	0.61	0.51	0.84	0.72	0.72	0.72	0.74
Hindi - HindiBERT	0.63	0.62	0.64	0.74	0.54	0.59	0.51	0.83	0.70	0.71	0.69	0.71
Multilingual BERT												
English - mBERT	0.70	0.71	0.70	0.79	0.61	0.63	0.59	0.85	0.77	0.78	0.77	0.79
Spanish - mBERT	0.62	0.65	0.61	0.77	0.57	0.59	0.55	0.84	0.74	0.74	0.74	0.75
French - mBERT	0.68	0.68	0.69	0.77	0.58	0.59	0.56	0.84	0.72	0.72	0.72	0.73
Portuguese - mBERT	0.66	0.67	0.67	0.77	0.54	0.55	0.53	0.83	0.71	0.71	0.71	0.72
Chinese - mBERT	0.62	0.61	0.64	0.74	0.54	0.60	0.49	0.84	0.69	0.70	0.69	0.71
Hindi - mBERT	0.47	0.48	0.47	0.66	0.43	0.54	0.35	0.82	0.64	0.65	0.64	0.67

Table 3: Disparity between English and non-English languages using monolingual and multilingual models. Performance of the task and language-specific text-only classification models on 3 datasets and 6 languages.

gingFace). For monolingual models, we refer to the leaderboard maintained by Nozza, Bianchi, and Hovy (2020) and select the best performing models for a specific language. Namely, we select BETO for modeling Spanish text (Cañete et al. 2020), CamemBERT for French (Martin et al. 2020), BERTimbau for Portuguese (Souza, Nogueira, and Lotufo 2020), ChineseBERT for Chinese (Cui et al. 2020), and HindiBERT for Hindi (Doiron 2020). We adopt the same model training and hyper-parameter selection strategies as for the English language models discussed above. Training a classification model for each of the five non-English languages across the three tasks gives us a total of 30 non-English text classification models. Our training strategies allow us to compare the *best* text classification models for all the languages for each of the three tasks individually.

Fine-tuned text representations: Once fine-tuned, the text classifiers can be used to extract representations for any input text by taking the output of the penultimate layers. These representations, also called *embeddings*, capture attributes of the text that the model has learned to use for categorizing the input into the target classes, and therefore can be fed to the multimodal classifier as a representation of the text part of the multimodal input. We obtain this latent representation of input text, denoted by vector $\mathbf{T}$ (with dimension 768), by averaging the token-level outputs from the penultimate layer of the fine-tuned classification models.

Evaluation metrics: We compute standard classification metrics to evaluate the performance these text-only classifiers on the test sets of respective datasets. Since crisis humanitarian post detection and emotion detection are multi-class classification tasks, we compute macro averages of class-wise F_1, precision, and recall scores along with the overall classification accuracy. However, since fake news detection is a binary classification task, we compute the F_1, precision, and recall scores for the positive class (i.e., `label:1 = fake`). Table 3 summarizes the performance of the text-only classifiers discussed above. Since the performance of deep learning models, especially BERT-based large language models, can possibly change with initialization schemes (Sellam et al. 2021), we vary the random initialization across different runs of the models and report the averages from 10 different runs.

Performance on English vs. non-English languages: In Table 3, we observe that the performance of text-only classification models is higher when the input is in the English language when compared against the performance of models that take other high-resource non-English languages as input. This trend is consistent across *(i)* both monolingual and multilingual models, *(ii)* the three tasks considered in this work as well as *(iii)* across all the classification metrics. For monolingual and multilingual models, the gap in performance on English and non-English languages varies with the task at hand as well as the non-English language being considered. For instance, for the crisis humanitarianism task with monolingual models, the drop in F_1 score of Spanish with respect to that of English is 9.5%, while it is 5.1% for the emotion detection task. For the same task, e.g., emotion detection, using monolingual models leads to performance drops that vary from 5.1% for Spanish to 11.4% for Hindi. It is noteworthy that the performance on non-English languages relative to each other maintains a near-uniform pattern across the three tasks for both monolingual and multilingual models – the performance is consistently the worst for Hindi; the performance on Chinese and Portuguese is relatively better, and the performance on Spanish and French is best when compared against other non-English languages. We revisit this observation and its potential causes in the Discussion section. In sum, our results indicate a language disparity exists due to the use of large language models in varied classification tasks — whether monolingual or multilingual. We recall that the adopted methodology – fine-tuning of pre-trained language models – is representative of the state-of-the-art NLP techniques that are frequently adopted for solving classification tasks (Li et al. 2020).

Benefits of Multimodal Learning

In this section, we focus on RQ2: can we leverage images with the help of multimodal learning to overcome the disparity between English and non-English languages.

Input	Crisis Humanitarianism				Fake News Detection				Emotion Detection			
	F_1	Precision	Recall	Accuracy	F_1	Precision	Recall	Accuracy	F_1	Precision	Recall	Accuracy
Image-only	0.42	0.45	0.42	0.52	0.15	0.54	0.09	0.81	0.94	0.94	0.94	0.95
Monolingual BERTs												
English + Image	0.73	0.74	0.72	0.82	0.60	0.63	0.58	0.85	0.85	0.87	0.84	0.86
Spanish + Image	0.72	0.73	0.71	0.82	0.59	0.63	0.57	0.85	0.82	0.83	0.81	0.82
French + Image	0.71	0.72	0.69	0.81	0.58	0.61	0.55	0.84	0.81	0.82	0.81	0.82
Portuguese + Image	0.71	0.71	0.70	0.80	0.59	0.60	0.58	0.84	0.81	0.82	0.81	0.82
Chinese + Image	0.70	0.69	0.70	0.80	0.58	0.62	0.54	0.84	0.80	0.80	0.79	0.81
Hindi + Image	0.68	0.69	0.67	0.80	0.56	0.61	0.51	0.84	0.78	0.79	0.77	0.80
Multilingual BERT												
English + Image	0.75	0.78	0.73	0.82	0.61	0.63	0.60	0.85	0.80	0.82	0.79	0.82
Spanish + Image	0.75	0.77	0.74	0.81	0.60	0.64	0.56	0.85	0.76	0.80	0.75	0.76
French + Image	0.74	0.84	0.71	0.83	0.58	0.60	0.57	0.84	0.76	0.76	0.76	0.77
Portuguese + Image	0.76	0.76	0.76	0.82	0.56	0.55	0.57	0.83	0.77	0.77	0.78	0.79
Chinese + Image	0.73	0.75	0.71	0.80	0.55	0.52	0.57	0.83	0.77	0.79	0.76	0.79
Hindi + Image	0.64	0.68	0.61	0.78	0.46	0.57	0.38	0.83	0.75	0.76	0.74	0.76

Table 4: Image-only and multimodal classfication performance. Performance of task-specific image-only classifiers (Row 1) and task- and language-specific multimodal classifiers (both monolingual and multilingual).

Image-Only classification model: To investigate the predictive power of images without textual information, we develop and evaluate image-only classifiers for each dataset. Similar to text classifiers, we apply a fine-tuning approach to train the task-specific image classifiers. We first freeze the weights of VGG-16 (Simonyan and Zisserman 2015), a popular deep Convolutional Neural Network, pre-trained on ImageNet (Deng et al. 2009), a large-scale generic image classification dataset. Then, we swap the last layer from the original model to three fully connected hidden layers with dimensions `4096`, `256`, and `num-of-classes`, respectively. Finally, retrain these three layers to adapt the image distribution in each dataset.

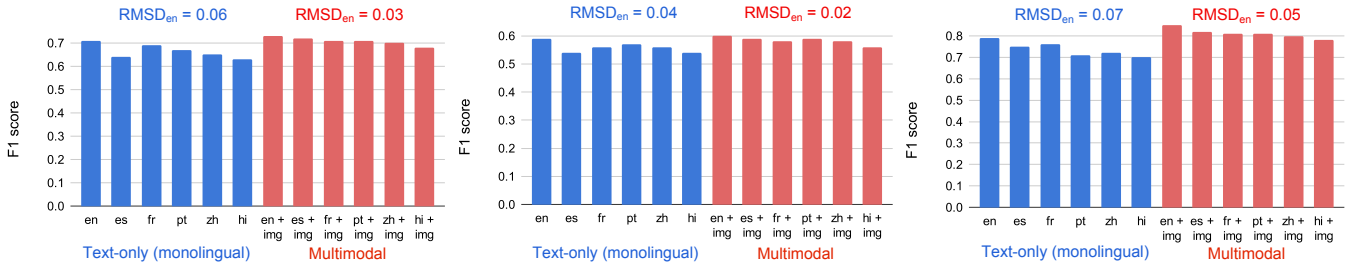
As images in our datasets have various dimensions, we apply a standard image pre-processing pipeline so that they can fit the pre-trained VGG-16 model's input requirement. We first resize the image so that its shorter dimension is 224, then we crop the square region in the center and normalize the square image with the mean and standard deviation of the ImageNet images (Deng et al. 2009).

To train and evaluate image-specific classifiers, we use the same splits in text-only models to divide images into the train, validation, and test sets. We use Adam optimizer (Kingma and Ba 2014) with a learning rate of 10^{-4} for each dataset. To avoid overfitting, we use early stopping to stop training when the loss value on the validation set stops to improve for 10 consecutive epochs. Finally, we extract the image embeddings, denoted by $\mathbf{I}$, from image-specific classifiers by computing the neuron activations from the penultimate layer (with dimension 256) as a latent representation of the image information for our multimodal models.

Multimodal classification model: We implement a multimodal classifier (Ngiam et al. 2011) that fuses the latent representations of individual modalities (text and image) to perform classification based on the joint modeling of both input modalities. We feed the concatenation of fine-tuned text and image representations (i.e., $\mathbf{T} \oplus \mathbf{I}$) to the multimodal classifier, which is essentially a series of fully connected layers with ReLU activation (Agarap 2018). The architecture of the multimodal classifier comprises an input layer (1024 neurons), 3 hidden layers (512, 128, 32 neurons), and an output layer (neurons = number of classes in the dataset). We train a multimodal classifier for each language in each task. Similar to image-only and text-only classification models discussed above, for each training instance, we use Adam optimizer (Kingma and Ba 2014) with a learning rate initialized at 10^{-4}. We use early stopping based on the validation set loss to stop the training and avoid overfitting on the train set.

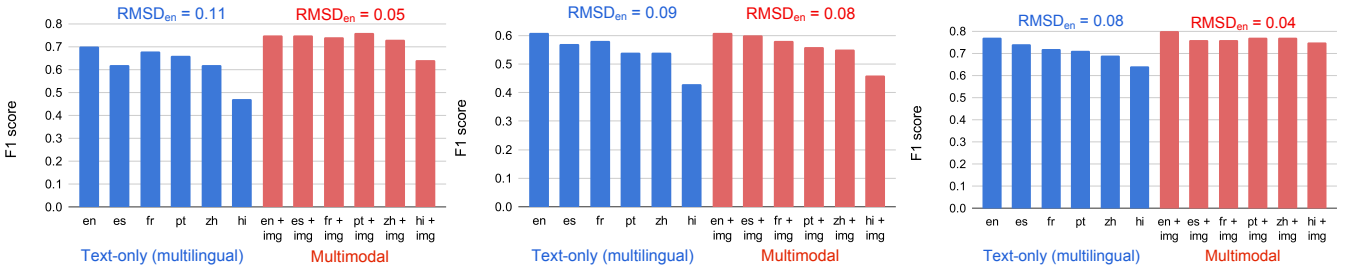
We use the same evaluation metrics to evaluate the image-only and multimodal classifiers as we did for the text-only ones, and report the average of 10 different runs in Table 4. Additionally, in Figures 4 and 5 we present the root-mean-squared deviation (RMSD$_{en}$) values of F_1 scores of non-English languages with respect to that of the English language for text-only and multimodal classifiers.

Multimodal learning boosts classification performance: As Table 4 shows, the classification performance for all the languages (English as well as non-English) improves considerably with the inclusion of images as an additional modality when compared against the performance of corresponding text-only classification models. This trend is consistent across all three datasets and both the set of models considered in our study — monolingual as well as multilingual. It is interesting to note that the benefit of including images, as indicated by the increase in performance metrics, is largely dependent on the *informativeness* of the images towards the classification task. For instance, for fake news detection, the image-only classifier achieves an F_1 score of 0.15, indicating poor distinguishability between real and fake news solely based on images in a news article. Consequently, the increase in the performance of the multimodal classifier over that of the monolingual text-only classifier is relatively marginal, ranging from 1.5% (F_1 increases from 0.59 to 0.60 for English) to 3.7% (F_1 increases from 0.54 to 0.56 for Hindi). In contrast, for the emotion detection task, the image-only classifier achieves an F_1 score of 0.94,



(a) Crisis Humanitarianism (b) Fake News Detection (c) Emotion Detection

Figure 4: Comparing F_1 scores on non-English and English text for both text-only and multimodal classifiers using *monolingual* language models. RMSD$_{en}$ denotes the root-mean-square deviation of the F_1 scores achieved by non-English classifiers with respect to the that of the corresponding English classifier. The RMSD$_{en}$ values for multimodal models are lower than those for monolingual text-only models.



(a) Crisis Humanitarianism (b) Fake News Detection (c) Emotion Detection

Figure 5: Comparing F_1 scores on non-English and English text for both text-only and multimodal classifiers using *multilingual* models. RMSD$_{en}$ denotes the root-mean-square deviation of the F_1 scores achieved by non-English classifiers with respect to the that of the corresponding English classifier. The RMSD$_{en}$ values for multimodal models are lower than those for multilingual text-only models.

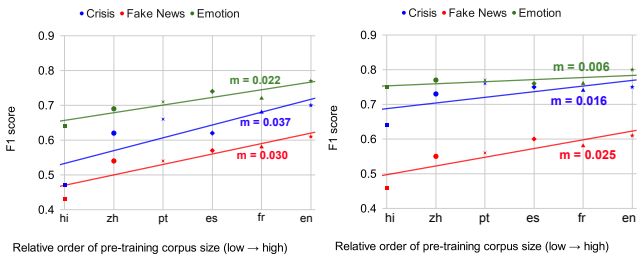
indicating extremely good distinguishability between emotion categories solely based on images. As a consequence, the increase in the performance of the multimodal classifier over that of the monolingual text-only classifier ranges from 7.6% (F_1 increases from 0.79 to 0.85 for English) to 11.4% (F_1 increases from 0.70 to 0.78 for Hindi). We observe the same trends for multilingual models as well.

Multimodal learning helps in bridging the gap between English and non-English languages: The results discussed so far indicate: *(i)* the performance of the state-of-the-art techniques for non-English languages is worse than the performance of the state-of-the-art techniques for the English language, and *(ii)* incorporating images as an additional modality using multimodal learning leads to better classification performance when compared against the performance of text-only counterparts. However, a crucial question remains to be answered: can multimodal learning help in overcoming the language disparity between English and non-English languages? To answer this, we focus on the root-mean-square deviation (RMSD$_{en}$) scores presented in Figures 4 and 5. RMSD$_{en}$ is calculated by taking the root of the average of the squared pairwise differences between F_1 scores for English and other non-English languages. We compute the RMSD$_{en}$ scores for both monolingual and multilingual models. It is clear that the RMSD$_{en}$ of F_1 scores achieved by non-English classifiers with respect to the F_1 score achieved by the English classifier are lesser with multimodal input when compared against text-only input. For monolingual models, the drops in RMSD$_{en}$ values are 50.0% ($0.06 \rightarrow 0.03$; Figure 4(a)), 50.0% ($0.04 \rightarrow 0.02$; Figure 4(b)), and 28.6% ($0.07 \rightarrow 0.05$; Figure 4(c)) for crisis humanitarianism, fake news detection, and emotion detection, respectively. Similarly, for the multilingual models, the drops in RMSD$_{en}$ values are 54.5% ($0.11 \rightarrow 0.05$; Figure 5(a)), 11.1% ($0.09 \rightarrow 0.08$; Figure 5(b)), and 50.0% ($0.08 \rightarrow 0.04$; Figure 5(c)) for crisis humanitarianism, fake news detection, and emotion detection, respectively. The drop in deviation with respect to the scores for English demonstrates that images are effective in bridging the gap between English and non-English languages. This is also pictorially depicted in Figures 4 and 5, as the red bars (with multimodal input) are more uniform in length than the blue bars (with text-only input).

Results on human-translated test set: To evaluate the performance of trained models on a sample that is free from the noise introduced by automated translators, we evaluate all the trained models for the crisis humanitarian task on the human-translated subset of the test set. Table 5 reinforces our observations — the disparity between English and non-English languages exists due to both monolingual and mul-

Language & Model	Crisis Humanitarianism	
	Language-only F_1 score	Multimodal F_1 score
Monolingual BERTs		
English	0.68	0.72
Spanish	0.63	0.69
French	0.64	0.70
Portuguese	0.63	0.68
Chinese	0.64	0.67
Hindi	0.61	0.66
Multilingual BERT		
English	0.69	0.73
Spanish	0.62	0.72
French	0.63	0.72
Portuguese	0.61	0.69
Chinese	0.60	0.66
Hindi	0.44	0.61

Table 5: Classification performance on *human* translated crisis humanitarianism test set. Performance of language-only and multimodal classifiers. The reported values are averages of 10 different runs.



(a) Text-only models (b) Multimodal models

Figure 6: Relation between pre-training corpus size and classification performance. The languages on the x-axes are ordered as per their representation in the pre-training corpora of mBERT; y-axes report the F_1 scores achieved on all the considered languages and task after task-specific fine-tuning. m denotes the slope of the task-wise trend lines.

tilingual language models and multimodal learning helps in reducing this performance gap. For monolingual and multilingual models, the RMSD$_{en}$ values drop from 0.05 to 0.04 and from 0.15 to 0.06, respectively.

Discussion

Our study demonstrates that in the context of societal tasks – as demonstrated by our focus on three datasets – the performance of large language models on non-English text is subpar when compared to the performance on English text. In the subsequent discussion, we highlight how this could have possibly threatening implications on the lives of many individuals who belong to underserved communities.

Furthermore, we empirically demonstrate that using images as an additional modality leads to a lesser difference between the performance on English and non-English text, as indicated by decreased RMSD$_{en}$ values. While existing studies have focused on developing advanced monolingual language models that can boost the performance on specific non-English languages to bridge the performance gap, we demonstrate the benefits of including other complementary modalities, especially those that are language-agnostic. Decreased RMSD$_{en}$ values indicate that if images are considered along with the text, the performance on all languages is not only better than when *only text* is considered, but it is also comparable across English and non-English languages.

Implications of language disparity with text-only models: In the context of social computing, disparities between English and non-English languages can lead to inequitable outcomes. For instance, as per our observations, if state-of-the-art NLP techniques that are centered around BERT-based language models are adopted to detect humanitarian information during crises, the detection abilities would be poorer for social media posts in non-English languages than those in English, causing delayed interventions. In countries like the United States, where non-English languages like Spanish and Chinese are spoken by a considerable number of people (AAAS 2016), this disparity could exacerbate the effects of discrimination and prejudice that they already face (PewResearch 2018). Similarly, poor emotion recognition in specific non-English languages can lead to unhelpful or even harmful outcomes in scenarios where the output of emotion recognition informs mental health interventions. Furthermore, poor fake news detection in specific non-English languages can lead to lacking correction and mitigation efforts, leading to relatively worse outcomes for non-English speaking populations.

Implications of reduced language disparity with multimodal models: People use multiple content modalities – images, text, videos, and audio clips, to share updates on social platforms. Visual modalities (like images and videos) transcend languages and are extremely informative in scenarios like crisis information detection and emotion detection. Combining our multimodal approach with existing text-only approaches for better modeling of non-English text can present *complementary gains*, leading to a reduced gap between English and non-English languages. In other words, an approach that complements existing approaches that focus on only text can be expected to provide gains even as the language-specific text-only approaches improve and evolve.

Dependence of performance on pre-training corpus size: The multilingual language model used in this study — mBERT, was pre-trained on huge corpora using self-supervised objectives (Devlin et al. 2018). The data sizes (in GiB) in mBERT's pre-training corpus have the relative order `en` > `fr` > `es` > `pt` > `zh` > `hi` (Conneau et al. 2020). As shown in Figure 6(a), the relationship between the language-specific corpus size that mBERT is trained on and the classification performance obtained after task-specific fine-tuning, is clear: larger representation in the pre-training corpus is related to better performance on downstream tasks. This trend reinforces the findings of Wu and Dredze (2020) in our context — the performance of large language models drops significantly as the considered languages have less pre-training data. This is concerning because, as Bender et al. (2021) argue, "the large, uncurated, and Internet-based datasets" that these language models are trained on "encode the dominant/hegemonic view, which further harms people at the mar-

gins." However, as shown in Figure 6(b), incorporating images using multimodal learning leads to a weakened dependence on the pre-training corpus size. This is indicated by the reduced slopes (m) of the trend lines across all three tasks. In effect, we demonstrate that multimodal learning, if adopted in the fine-tuning stages of approaches that employ large language models, could help in overcoming the well-recognized dependence of downstream performance on language-specific pre-training corpus size.

Beyond the theoretical implications discussed above, we believe our methods demonstrate the crucial role that multimodal learning can play in the equitable dissemination of NLP-based services to a broader range of the global population. The systems that make inferences based on text data alone can be adapted to include the information contained in images, wherever possible, leading to better detection abilities on the non-English text and thereby bridging the gap between English and non-English languages. As our evaluation on human-translated and machine-translated text demonstrates, our proposed approach is compatible with setups that infer information directly from non-English text and with the approaches that first translate non-English text to English and then infer information from the translations.

Limitations and future work: Large language models such as T5 and their corresponding multilingual variants mT5 overcome several limitations of BERT and mBERT by adopting different pre-training strategies. We specifically focused on BERT-based language models as representatives of large language models – note that our study aimed to understand the effectiveness of multimodal learning in overcoming the language disparity and not the relative performance of different language models. Since the underlying idea of fusing image and text representations can be applied to other language models as well, we believe that our insights and takeaways will also generalize to them.

In the future, we intend to experiment with low-resource languages to expand our claims to a wider set of languages. There are two major challenges on those fronts: *(i)* availability of parallel data, and *(ii)* identifying and developing the state-of-the-art text-only classification approaches for low-resource languages. A translation-based data creation pipeline will not work for low-resource languages and hence we may either curate the data by recruiting native speakers to translate the original examples from English or by collecting real data from social media for different languages. Furthermore, since the state-of-the-art classification approach for low resource languages may not be based on large language models (Wu and Dredze 2020; Nozza, Bianchi, and Hovy 2020), we intend to identify and develop those language-specific approaches.

Lastly, the current study focuses on bridging the gap that exists in classification tasks. As part of future work, we intend to explore other types of tasks that are relevant to the social computing theme. Such tasks include, analyzing the lifestyle choices of social media users (Islam and Goldwasser 2021) and context-based quotation recommendation (MacLaughlin et al. 2021). By including other modalities like images, these approaches may be extended to non-English speaking populations. However, while images are not bound by languages, their production and perception are culturally influenced (Hong et al. 2003). This cultural influence is more prominent in user-generated content that is abundant on social platforms (Shen, Wilson, and Mihalcea 2019). Therefore, it is important to consider the cultural confounds in the production and consumption of images while using them to train and infer from machine learning models.

Broader perspective, ethics, and competing interests: Developing powerful, accessible, and equitable resources for modeling non-English languages remains an open challenge. Our work argues that including information from other modalities, specifically images, can present new avenues to progress research in this direction. We believe this work will positively impact society by motivating researchers and practitioners to develop more reliable classifiers for non-English languages with applications to societal tasks. That said, it is worth noting that since images alone do not represent the entire cultural context, modeling techniques for non-English languages should continue to develop. Incorporation of new modalities alongside text also comes with additional challenges — for instance, the biases that computer vision models encode (Hendricks et al. 2018) need to be taken into consideration, and methods need to be developed to model cultural shifts in meaning for similar images (Liu et al. 2021). The authors involved in this study do not have any competing interests that could have influenced any part of the conduct of this research.

Conclusion

In sum, we have demonstrated that the adoption of large language models for building approaches for tasks aimed at detecting humanitarian information, fake news, and emotion leads to systematically lower performance on non-English languages when compared to the performance on English. We discussed how such a disparity could lead to inequitable outcomes. Furthermore, we empirically show that including images via multimodal learning bridges this performance gap. Our experiments yield consistent insights on 3 different datasets and 5 non-English languages, indicating their generalizability. We also discussed the reliance of large language models on pre-training corpus size and how adopting multimodal learning during fine-tuning stages can weaken this dependence, leading to a more consistent performance across all languages under consideration.

Acknowledgements

This research has been supported in part by NSF IIS-2027689, NSF ITE-2137724, Microsoft AI for Health, and IDEaS at Georgia Tech. We thank Sindhu Kiranmai Ernala, Sandeep Soni, and Talayeh Aledavood for helpful discussions in the early stages of the project. We acknowledge Shivangi Singhal (IIIT-Delhi, India) for providing us with the pre-processed multimodal fake news dataset. We also thank Bing He and Kartik Sharma for helping with translations, the CLAWS research group members for preliminary manual inspections of the translated text, and the anonymous reviewers for their constructive feedback.

References

AAAS. 2016. The state of languages in the US: A statistical portrait. https://www.amacad.org/publication/state-languages-us-statistical-portrait. Accessed: 2022-01-09.

Agarap, A. F. 2018. Deep learning using rectified linear units (ReLU). *arXiv:1803.08375*.

Alam, F.; Ofli, F.; and Imran, M. 2018. Crisismmd: Multimodal twitter datasets from natural disasters. In *AAAI ICWSM*.

Arviv, E.; Hanouna, S.; and Tsur, O. 2021. It'sa Thin Line Between Love and Hate: Using the Echo in Modeling Dynamics of Racist Online Communities. In *AAAI ICWSM*.

Basta, C.; Costa-jussà, M. R.; and Casas, N. 2019. Evaluating the Underlying Gender Bias in Contextualized Word Embeddings. In *Proc. of the Workshop on Gender Bias in NLP*.

Bender, E. 2019. The #BenderRule: On Naming the Languages We Study and Why It Matters. *The Gradient*.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *ACM FAccT*.

Caruana, R.; Lawrence, S.; and Giles, C. 2000. Overfitting in Neural Nets: Backpropagation, Conjugate Gradient, and Early Stopping. *NeurIPS*.

Cañete, J.; Chaperon, G.; Fuentes, R.; Ho, J.-H.; Kang, H.; and Pérez, J. 2020. Spanish Pre-Trained BERT Model and Evaluation Data. In *PML4DC at ICLR 2020*.

Choi, M.; Budak, C.; Romero, D. M.; and Jurgens, D. 2021. More than Meets the Tie: Examining the Role of Interpersonal Relationships in Social Networks. In *AAAI ICWSM*.

Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, É.; Ott, M.; Zettlemoyer, L.; and Stoyanov, V. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *ACL*.

Crook, B.; Glowacki, E. M.; Suran, M.; K. Harris, J.; and Bernhardt, J. M. 2016. Content analysis of a live CDC Twitter chat during the 2014 Ebola outbreak. *Comm'n. Res. Reports*.

Cui, Y.; Che, W.; Liu, T.; Qin, B.; Wang, S.; and Hu, G. 2020. Revisiting Pre-Trained Models for Chinese Natural Language Processing. In *EMNLP (Findings)*.

De Choudhury, M.; Counts, S.; and Horvitz, E. 2013. Social media as a measurement tool of depression in populations. In *ACM WebSci*.

De Choudhury, M.; Gamon, M.; Counts, S.; and Horvitz, E. 2013. Predicting depression via social media. In *AAAI ICWSM*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *IEEE CVPR*.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR*, abs/1810.04805.

Doiron, N. 2020. Hindi BERT on HuggingFace. https://huggingface.co/monsoon-nlp/hindi-bert. Accessed: 2022-01-09.

Duong, C. T.; Lebret, R.; and Aberer, K. 2017. Multimodal classification for analysing social media. *arXiv:1708.02099*.

Glasgow, K.; Fink, C.; and Boyd-Graber, J. 2014. " Our Grief is Unspeakable": Automatically Measuring the Community Impact of a Tragedy. In *AAAI ICWSM*.

Hedderich, M. A.; Lange, L.; Adel, H.; Strötgen, J.; and Klakow, D. 2021. A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios. In *NAACL-HLT*.

Hendricks, L. A.; Burns, K.; Saenko, K.; Darrell, T.; and Rohrbach, A. 2018. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Higgins, J.; LaSalle, A.; Zhaoxing, P.; Kasten, M.; Bing, K.; Ridzon, S.; and Witten, T. 2009. Validation of photographic food records in children: are pictures really worth a thousand words? *Euro. J. of Clinical Nutrition*.

Hong, Y.-Y.; Benet-Martinez, V.; Chiu, C.-Y.; and Morris, M. W. 2003. Boundaries of cultural influence: Construct activation as a mechanism for cultural differences in social perception. *J. of Cross-Cultural Psych.*

Houston, J. B.; Hawthorne, J.; Perreault, M. F.; Park, E. H.; Goldstein Hode, M.; Halliwell, M. R.; Turner McGowen, S. E.; Davis, R.; Vaid, S.; McElderry, J. A.; et al. 2015. Social media and disasters: a functional framework for social media use in disaster planning, response, and research. *Disasters*.

Islam, T.; and Goldwasser, D. 2021. Analysis of Twitter Users' Lifestyle Choices using Joint Embedding Model. In *AAAI ICWSM*.

Joshi, P.; Santy, S.; Budhiraja, A.; Bali, K.; and Choudhury, M. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *ACL*.

Junczys-Dowmunt, M.; Grundkiewicz, R.; Dwojak, T.; Hoang, H.; Heafield, K.; Neckermann, T.; Seide, F.; Germann, U.; Fikri Aji, A.; Bogoychev, N.; Martins, A. F. T.; and Birch, A. 2018. Marian: Fast Neural Machine Translation in C++. In *ACL 2018, System Demonstrations*.

Kern, M. L.; Park, G.; Eichstaedt, J. C.; Schwartz, H. A.; Sap, M.; Smith, L. K.; and Ungar, L. H. 2016. Gaining insights from social media language: Methodologies and challenges. *Psych. Methods*.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv:1412.6980*.

Lazer, D. M.; Baum, M. A.; Benkler, Y.; Berinsky, A. J.; Greenhill, K. M.; Menczer, F.; Metzger, M. J.; Nyhan, B.; Pennycook, G.; Rothschild, D.; et al. 2018. The science of fake news. *Science*.

Li, Q.; Peng, H.; Li, J.; Xia, C.; Yang, R.; Sun, L.; Yu, P. S.; and He, L. 2020. A survey on text classification: From shallow to deep learning. *arXiv:2008.00364*.

Liu, F.; Bugliarello, E.; Ponti, E. M.; Reddy, S.; Collier, N.; and Elliott, D. 2021. Visually Grounded Reasoning across Languages and Cultures. In *EMNLP*.

MacLaughlin, A.; Chen, T.; Ayan, B. K.; and Roth, D. 2021. Context-based quotation recommendation. In *AAAI*.

Martin, L.; Muller, B.; Ortiz Suárez, P. J.; Dupont, Y.; Romary, L.; de la Clergerie, É.; Seddah, D.; and Sagot, B. 2020. CamemBERT: a Tasty French Language Model. In *ACL*.

Microsoft. 2019. Neural Machine Translation Enabling Human Parity Innovations In the Cloud. https://www.microsoft.com/en-us/translator/blog/2019/06/17/neural-machine-translation-enabling-human-parity-innovations-in-the-cloud/. Accessed: 2022-01-09.

Mielke, S. J. 2016. Language diversity in ACL 2004 - 2016. https://sjmielke.com/acl-language-diversity.htm. Accessed: 2022-01-09.

Mocanu, D.; Baronchelli, A.; Perra, N.; Gonçalves, B.; Zhang, Q.; and Vespignani, A. 2013. The twitter of babel: Mapping world languages through microblogging platforms. *PloS One*.

Muller, B.; Anastasopoulos, A.; Sagot, B.; and Seddah, D. 2021. When Being Unseen from mBERT is just the Beginning: Handling New Languages With Multilingual Language Models. In *NAACL-HLT*.

Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; and Ng, A. Y. 2011. Multimodal deep learning. In *ICML*.

Nozza, D.; Bianchi, F.; and Hovy, D. 2020. What the [mask]? making sense of language-specific BERT models. *arXiv:2003.02912*.

Ofli, F.; Alam, F.; and Imran, M. 2020. Analysis of Social Media Data using Multimodal Deep Learning for Disaster Response. *arXiv:2004.11838*.

Pennebaker, J. W.; Francis, M. E.; and Booth, R. J. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*.

PewResearch. 2018. Latinos and discrimination. https://www.pewresearch.org/hispanic/2018/10/25/latinos-and-discrimination/. Accessed: 2021-09-10.

Pires, T.; Schlinger, E.; and Garrette, D. 2019. How Multilingual is Multilingual BERT? In *ACL*.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *JMLR*.

Rasmy, L.; Xiang, Y.; Xie, Z.; Tao, C.; and Zhi, D. 2021. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digital Med.*

Reynolds, L.; and Johnson, R. 2011. Is a picture is worth a thousand words? Creating effective questionnaires with pictures. *Practical Assessment, Research, and Eval.*

Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv:1910.01108*.

Sellam, T.; Yadlowsky, S.; Wei, J.; Saphra, N.; D'Amour, A.; Linzen, T.; Bastings, J.; Turc, I.; Eisenstein, J.; Das, D.; et al. 2021. The MultiBERTs: BERT Reproductions for Robustness Analysis. *arXiv:2106.16163*.

Shen, Y.; Wilson, S. R.; and Mihalcea, R. 2019. Measuring personal values in cross-cultural user-generated content. In *Int. Conf. on Social Informatics*. Springer.

Shu, K.; Mahudeswaran, D.; Wang, S.; Lee, D.; and Liu, H. 2018. FakeNewsNet: A Data Repository with News Content, Social Context and Dynamic Information for Studying Fake News on Social Media. *arXiv:1809.01286*.

Shu, K.; Sliva, A.; Wang, S.; Tang, J.; and Liu, H. 2017. Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD Explorations Newsletter*.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556*.

Singhal, S.; Kabra, A.; Sharma, M.; Shah, R. R.; Chakraborty, T.; and Kumaraguru, P. 2020. Spotfake+: A multimodal framework for fake news detection via transfer learning (student abstract). In *AAAI*.

Singhal, S.; Shah, R. R.; Chakraborty, T.; Kumaraguru, P.; and Satoh, S. 2019. Spotfake: A multi-modal framework for fake news detection. In *IEEE Int. Cont. on Multimedia Big Data (BigMM)*. IEEE.

Souza, F.; Nogueira, R.; and Lotufo, R. 2020. BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems BRACIS*.

Strubell, E.; Ganesh, A.; and McCallum, A. 2019. Energy and Policy Considerations for Deep Learning in NLP. In *ACL*.

Sun, C.; Huang, L.; and Qiu, X. 2019. Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence. In *NAACL-HLT*.

Teixeira, T.; Wedel, M.; and Pieters, R. 2012. Emotion-induced engagement in internet video advertisements. *J. of Marketing Res.*

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv:1910.03771*.

Wu, S.; and Dredze, M. 2020. Are All Languages Created Equal in Multilingual BERT? In *Proc. of the Workshop on Representation Learning for NLP*.

Xue, L.; Constant, N.; Roberts, A.; Kale, M.; Al-Rfou, R.; Siddhant, A.; Barua, A.; and Raffel, C. 2021. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *NAACL-HLT*.