Multilingual Social Media Text Generation and Evaluation with Few-Shot Prompting

Mack Blackburn
Leidos
mack.blackburn@leidos.com

Lizzie Yates
Leidos

Ning Yu
Leidos

Abstract

This work adapts large language models to generate multilingual social media text that meets several objectives simultaneously: topic relevance, author style consistency, and reply validity. Leveraging existing online information behavior simulators, which currently only forecast activities but not content, our approach comprised of generalizable prompt formation and efficient evaluation to produce a believable, personalized, and responsive synthetic social network. According to some preliminary experiments, our multi-objective prompt formation and automatic evaluation/selection methods are able to yield a significant number of high-quality synthetic texts according to both standardized and trained metrics.

1 Introduction

Our work on generation of synthetic social media text is motivated by existing technologies to simulate and forecast behavioral phenomena and information spread on social media platforms as part of the DARPA SocialSim program (Murić et al., 2020). While these simulators can forecast social media activity such as who will post on which topic or who will reply to whom at what time, they do not produce any text values for simulated activities. Our work fills this gap with text generation and provides a complete picture of simulated social network landscape. Novel methods for text generation are frequently explored; however our task involves the unusual aspect of generating convincing social media posts and replies in multiple languages for simulated online dialogue without human involvement and targets specific topics, author styles, and responses. In order to achieve these multiple objectives, traditional approaches like transfer learning (Raffel and Liu, 2020) make separate calls to large languages models; we leverage few-shot prompting to reduce such computing expensive calls and developed efficient automated evaluation metrics for synthetic text selection.

We employ commonly used text generation metrics (Sai et al., 2020) including BLEU scores and find that they only capture some aspects of what makes a text high-quality in our context. In general, most existing evaluation methods for synthetic text have limitations of some kind (Huang and Huang, 2020). Some synthetic texts can achieve metrics higher than real ground-truth text according to these standard metrics. Some newer metrics are too computational expensive and not suitable for supporting large scale text evaluation and selection. Therefore, we implemented three generalizable and runtime-efficient evaluation methods to measure generated text for their topic and author relevance as well as reply flow within a network, and also evaluated against more standard metrics.

Our social media text generation approach is efficient, language agnostic and generalizable, and can be used at scale to mimic social media networks with millions of simulated activities. The resulting synthetic information networks can support media analysts' training exercises, or provide large-scale datasets for AI/ML modeling studying online information behavior.

2 Related Work

There has been a significant amount of work on the use of prompts to generate desired types of text using language models such as GPT-NEO (Black et al., 2021). Multi-prompt learning incorporates multiple prompts, either answered or unanswered, into the text generation model prompting paradigm (Brown et al., 2020). Prompt augmentation uses multiple answered prompts (Liu et al., 2021). Our dataset

Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM), pages 417 - 427

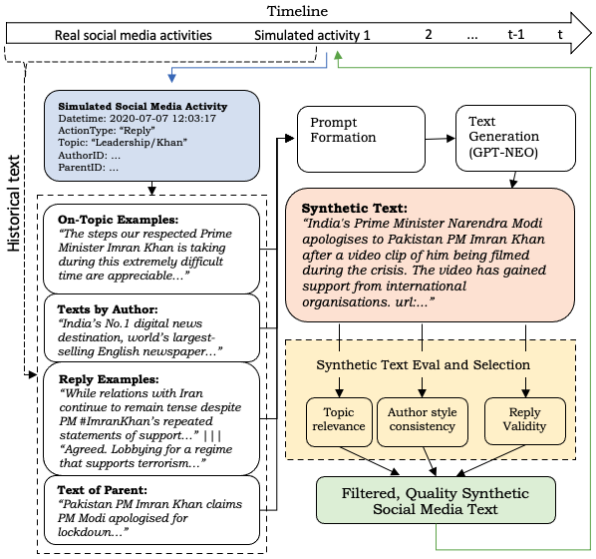

Figure 1: Text Generation and Evaluation Pipeline. In chronological order, each simulated activity is fed into this pipeline and the output is an enriched social media activity with generated text. Historical real and generated text are sampled for prompt formation.

contains many real world examples matching the multi-objective goals and many statement-reply pairs, so few-shot learning and prompt augmentation are good fits for our use case. While we are able to collect large volumes of social media text, manual annotation of text with specific labels is much more intensive, and few-shot learning allows us to utilize a smaller manually-annotated set.

3 Approach

,

3.1 Text Generation from Simulated Social Media Activities

Figure 1 depicts the high level pipeline. In chronological order, this pipeline sequentially generates and selects text for each simulated online activity. For any given simulated activity, the pipeline follows 4 steps:

1. Create a prompt that includes examples of real world historical texts (i.e., ground-truth) by the same author, followed by ground-truth texts on the same topic. Example texts were selected randomly. When the simulated activity is intended to be a reply, the prompt also includes examples of ground-truth statement-reply pairs, followed by the text of the parent text that the activity is meant to be responding to.

2. Feed formed prompts into a language model which returns generated text. We selected GPT-NEO because it has the best performance for text generation upon open sourced language models at the time when we were working on this project.

3. Evaluate and select generated text based on both standard and task-specific metrics to measure the three objectives: Topic relevance, authorship verification, and sentence-pair classification models (see 5).

4. Fill the simulated activities with generated text and move onto next simulated event on the timeline.

This pipeline for synthetic text evaluation and selection operates at scale and for multi-lingual/styled generation as well.

3.2 Prompt Formation

Single objective Few-Shot Prompts To generate text that is on-topic, in the style of a particular author, or responding to a particular statement, we use prompts that incorporate multiple examples of real-world text with these desired attributes. Specifically, on-topic prompts are selected from historical tweets based on their manual topic annotation described in section 5.2.3; user-focused prompts consist of real tweets by the same user with Twitter user's bio where available; Reply-focused prompts are more complicated and consist of examples of statement-response pairs as reference:

> *"The following is a list of tweet and response pairs:*
> *{{gold statement 1}} /// {{gold response 1}}*
> *===*
> *{{gold statement 2}} /// {{gold response 2}}*
> *===*
> *{{gold statement 3}} /// ..."*

The parent tweet will also be incorporated into the prompt for generation of a reply. Because we are generating text for large-scale simulations of social media, a simulated reply can be corresponding to either a real tweet or a simulated tweet. In the latter, we look

Language Code	Count
en	2,065,581
ur	504,485
hi	196,362
la	44,959
sw	21,410
ms	20,993
zh	17,722
id	15,853
mr	14,500
xh	12,292

Table 1: Count of each of the top ten languages as detected by the langid library

up the specific synthetic text for the parent tweet. Because of this, text generation occurs sequentially in the chronological order of the simulated activities.

Multi-Objective Few-Shot Prompts In addition to evaluations of text generated with a single objective, we also generate and evaluate texts with two objectives or three objectives at once in the case of synthetic replies. Multi-objective prompts are formatted by concatenating multiple single-objective prompts.

4 Dataset

4.1 Social Media Collection

For the DARPA SocialSim program, we collected online discussions relevant to the China-Pakistan Economic Corridor (CPEC) from multiple social media platforms. The primary platforms by data volume are Twitter and YouTube, and we will focus on Twitter for the rest of the paper. Due to the nature of the event, we create a list of keywords in English, Hindi, Urdu, Chinese and several other regional languages to query tweets and replies or YouTube video titles and description. Some keyword examples are: "china pakistan economic corridor", "cpec", "ਇਕ ਬੈਲਟ ਇਕ ਰੋਡ", and "ای کے بی ایل ٹ ون روڈ". The counts of each of the top ten languages in the dataset as detected by the langid python library (Lui and Baldwin, 2012) are shown in table 1. The top three languages by volume are English, Urdu, and Hindi.

Annotation From the collected social media data covering almost 5 million Tweets and YouTube comments, we selected roughly 5,000 of the most-interacted with texts to pass to three in-house manual annotators, who annotated for 21 distinct topic labels with a cross-label average Cohen's Kappa inter-annotator agreement of 0.78. Detailed annotation procedure can be found in (Blackburn et al.). As examples of topic labels, the label "benefits/development/jobs" refers to discussion of jobs brought by the CPEC program, and the label "controversies/china/border" refers to discussion of border disputes in the China-India-Pakistan region. The set of annotated texts were used to train a supervised text classifier (F1 score of 0.73 across all 21 topics). While the manually-annotated examples have been leveraged to provide few-shot prompt examples for text generation, the classifier was used to evaluate the topic relevance of generated text, see 5.2.3.

4.2 Simulated Social Media Activities

We use simulated social media activities generated by SocialCube (Tarek Abdelzaher, 2020) as a template for the time, author, and topic of synthetic social media activities, and fill in the text value with synthetic texts that fit the desired attributes. SocialCube takes real world social media activity and news event data in the training phase, and returns social media activity simulation for the following testing phase. We used one of the simulation results that contains 1,037,782 simulated activities for a total span of 27 days. After removing retweets that do not require unique text, 149,829 activities left require synthetic text values (e.g. new tweets and replies).

5 Evaluation and Metrics

Ground Truth In our tests, the ground truth used for evaluation contains real social media texts with specific attributes. There are three groups of ground truth: one group of texts manually annotated with topic labels (the test set of the manual annotation process in section 5.2.3), a group of texts by specific authors (the test set of the data used to train the authorship verification model in section 5.2.1), and a group of valid and invalid statement-response pairs (the test set of the dataset in section5.2.2).

5.1 Standardized Metrics

Median Sentence BLEU Sentence BLEU calculated with the sacrebleu library (Post, 2018) for generated message, treating the ground truth texts with the same class (i.e., topic or author/user) as reference. Median is calculated over distribution of scores. We choose median because mean can be effected by outliers, and we are concerned with the quantity of texts with high scores rather than overall average. (Papineni et al., 2002)

Self-BLEU Sentence BLEU calculated with the sacrebleu library for generated message, treating all texts with the same class and generation method as reference. Median is calculated over distribution of scores.(Zhu et al., 2018)

GEM Metrics Library We also run evaluation with multiple metrics from the GEM metrics library trying to covering different aspects of generated text: descriptive, diversity,lexical and semantic measurements. Some examples include BLEURT, Distinct-1 (the ratio of distinct unigrams over the total number of unigrams), Entropy-1 (the Shannon Entropy over unigrams), and text length metrics (Gehrmann et al., 2021).

5.2 Trained Metrics

Since none of the standardized metrics for text generation directly measure the three objectives of our use case, we also develop three specific evaluation methods respectively leveraging the ground truth data.

5.2.1 Authorship Verification for Author Style

Using ground-truth social media data, we identify all users with over 20 unique text-valued posts, and store those users' posts into train and test sets with the ratio 0.7 to 0.3. From the stored posts, we construct 50,000 training and 10,000 test pairs, where pairs of texts by the same user are in the "1" class and pairs of texts by two separate users are in the "0" class. We then finetune the distiluse-base-multilingual-cased language model on the pairs using contrastive loss with cosine for 5 epochs (Sanh et al., 2019). Training of the model uses the SentenceTransformers library with default parameters (Reimers and Gurevych, 2019). When using cosine similarity of the embeddings of the fine-tuned model as an indicator for authorship on the test set, we find a ROC-AUC of 0.94. The ROC-AUC of 0.94 is an evaluation of how well the cosine similarity metric distinguishes between texts by different authors in the test set.

We apply the author fine-tuned language model to each generated text, and measure the cosine similarity of the vector of each text to the centroid of the vectors for the ground truth texts by the intended author. The cosine similarity is used as an indicator of the degree to which the synthetic text matches the style of the intended author.

5.2.2 Statement-Response Pair Classification for Reply Detection

Data Preparation We train a sentence pair classifier to determine whether a response replies coherently to an original sentence. We begin by extracting around 350,000 real tweet-response pairs from our curated Twitter dataset. Roughly 15 percent of the data is set aside as holdout data, while the rest is processed for training. The training data is halved, with the first half being true pairs and receiving a label of 1. The second half of the training data is then halved again, and either shuffled row wise or column wise, providing "untrue" or 0-labeled tweet-response pairs while still maintaining contextual relevance. The pairs is then concatenated, shuffled, and set aside for training and evaluation.

Classification For the sentence-pair classification task, we utilize the Simple Transformers library (Rajapakse, 2020). Our final configuration is set up for a maximum sequence length of 512 to be trained over 2 epochs with an Adam optimizer and learning rate of 4e-5. Our trained model is evaluated on our evaluation dataset and returns a F1 score of .81, accuracy of .79, AUROC of .84 and AUPRC of .77. We test several model types including BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), distilbert (Sanh et al., 2019) and xlnet (Yang et al., 2019). BERT, and specifically bert-based-multilingual-cased, provides the strongest results for our use case.

Model	Precision	self-BLEU	Mean Topic Similarity	median BLEU
Ground Truth	0.91	38.83	0.64	38.83
gpt-neo-1.3B	0.35	67.03	0.55	43.07
gpt-neo-125M	0.33	57.59	0.46	39.18

Table 2: Topic-Relevance Evaluation Metrics, Across all Languages

5.2.3 Topic Classification Precision

We train a supervised topic classifier using the manually-annotated set described in section 4.1. Given a piece of generated text, if this classifier actually labels it with the intended topic, we assume the generated text is relevant to the intended label.

Topic Relevance In addition to the classifier precision metric, we also use the "distiluse-base-multilingual-cased" language model, which is appropriate for our multilingual dataset, and is specifically well-suited to measure semantic similarity with cosine (Cer et al., 2018). Vectors are extracted for each generated text and each ground-truth text. Ground-truth texts are grouped by topic class and an average vector for each class is computed. For each generated text, the cosine similarity of its vector to the ground-truth vector of the same class is measured. Cosine similarities are averaged across texts.

6 Evaluation and Discussion

6.1 Single-Objective Text Generation Evaluation

Topic Relevance Table 2 shows topic-relevance metrics for GPT-NEO 125M and 1.3B compared to the same metrics for real ground truth texts. For the ground truth, self-BLEU is the same as median BLEU, because median BLEU always uses the ground truth as references for computation. For some metrics such as BLEU there is an acceptable range rather than strictly "higher is better". The lower precision score of GPT-generated text compared to ground truth may show that generating text exactly on an exact topic is still a challenge.

Author Style As shown in table 3, GPT-NEO 1.3B with prompts incorporating the text of a user's self-reported social media bio and example texts by the user generated synthetic texts most similar to real ground truth texts by the same authors, according to our metric.

Method	Mean Author Similarity±Std Dev.
Ground Truth	0.93±0.08
GPT-NEO-1.3B + User Bio	0.86±0.14
GPT-NEO-125M	0.85±0.14
GPT-NEO-125M + User Bio	0.84±0.11
GPT-NEO-1.3B	0.82±0.13
Inverse Ground Truth	0.64±0.20

Table 3: Author Style Evaluation. "Mean author similarity" shows the mean of the cosine similarities between texts and the centroid of the vectors for the user's ground-truth texts. "Inverse Ground Truth" shows this metric computed on ground truth texts compared to texts by different authors, across all languages.

Synthetic Reply Evaluation Based on our evaluation shown in table 4, generating realistic synthetic replies is possible, but a significant portion of the synthetic texts generated may not be properly coherent or satisfactory replies.

Model	Mean Reply Score±Std Dev.
Ground Truth	0.89±0.31
GPT-NEO-125M	0.47±0.50
GPT-NEO-1.3B	0.41±0.49
Inverse Ground Truth	0.33±0.46

Table 4: Reply Evaluation. "Mean reply score" shows the averaged predictions of the reply classifier model. Ranged 0-1 where 0 is 0% valid replies and 1 is 100%. "Inverse Ground Truth" shows this metric computed on non-reply ground truth text pairs. Computed across all languages.

6.2 Multi-Objective Prompting Evaluation

As shown in table 5, the mean scores of most metrics degrade on the multi-objective task as compared to the single-objective tasks. However, we are still able to use our evaluation

Metric	GPT-NEO-1.3B	GPT-NEO-125M
Mean Topic Similarity ±Std Dev.	0.33±0.24	0.29±0.21
Mean Reply Score ±Std Dev.	0.37±0.48	0.39±0.49
Mean Author Similarity ±Std Dev.	0.76±0.14	0.76±0.18

Table 5: Evaluation of Texts from Multi-Objective Prompt, Across all Languages.

metrics as a filter to separate higher quality texts from lower quality ones, and still generate a relatively large number of high-quality synthetic texts, as described in section 6.3. Texts with a topic relevance score above 0.6, a user similarity score above 0.7, and a reply validity score above 0.6 are marked, and stored for use as text values of the simulated social media activities.

GEM Metrics Evaluation We use several metrics from the GEM metrics repository (Gehrmann et al., 2021) to measure the generated text from more perspectives. Some results are shown in table 6. Looking at some of the descriptive metrics, mean text length for GT and language models are similar. This is not surprising given the character limit for tweets. However, the range of generated tweet length varies much more than ground truth: as short as 1 and as long as 60+. Text length limit information can be introduced in the future to avoid generating text longer than allowed. Not included in the result table, the vocabulary size of ground truth, for topic or user relevance, is always much smaller than generated text. This suggests that machine generated text may contain terms that are not used often for social media. Conducting domain adaptation on the pre-trained language model to make it more relevant to social media data may help reduce this difference. Looking at some of the diversity metrics, the Distinct-1 metrics indicates ground truth text is much more diverse than generated text. This could due to the high creativity of language expression in social media. When it comes to semantic metrics, even ground truth tweets achieve pretty low BLUERT score. Generated text is worse, with bigger GTP-neo model performs slightly better. BLUERT is calculated for a small set of evaluation set due to its high computing demand, and we will look at this again when calculate it against the complete evaluation set.

Multi-Lingual Evaluation Because our dataset is multilingual, we also report the GEM metrics across the top 5 languages in our dataset: English, Farsi, Hindi, Urdu, and Chinese in table 7. Additional multilingual metrics are reported in appendices A, B, and C. English has the best performance in terms of BLEURT, followed by Hindi and Chinese. This could due to lack of real world data in our collection for other languages to create good prompts or the lack of explicit multilingual capabilities of GPT-NEO. All other evaluations reported in this paper are computed on texts regardless of language, including English and others.

6.3 Scalability and Runtime

We use a single-GPU instance on AWS (16vCPUs, 1 GPU, 64GB Memory) for generation and evaluation. The single-GPU machine was able to generate 29,150 synthetic texts per hour in total using GPT-NEO 125M, but after evaluation and selection of higher-quality texts, that number comes to 5,658 high-quality synthetic texts per hour.

7 Limitation

Due to the shortage of time and computing resource, we didn't finish running some of the heavy metrics such as NUBIA(Kane et al., 2020) for measuring faithfulness and BERTScore(Zhang* et al., 2020) for better semantic measurement. We also didn't measure the entire pipeline, e.g., comparing generated text and real text in the same simulation time frame and measuring the impact of generated and filtered text in simulation. We will address some of these limitations in our next steps. We will also try to evaluate the filtering step with manual review of filtered texts, and by testing the impact of synthetic texts on social media simulations.

8 Conclusions and Future Work

We show that it is feasible to generate and evaluate synthetic social media texts which not only focus on a desired topic, but also mimic an author style and properly respond to

Model	Mean Topic BLEURT	Mean User BLEURT	Distinct-1	Entropy-1	Mean Text Length
gpt-neo-1.3B	0.15±0.08	0.15±0.1	0.61±0.11	6.62±0.48	23.38±19.48
gpt-neo-125M	0.15±0.08	0.15±0.09	0.59±0.13	6.32±0.63	23.51±19.62
GT	0.21±0.06	0.25±0.09	0.85±0.12	4.59±0.85	27.03±1.39

Table 6: Selected GEM Metrics Across all Languages

Model	Lang	Mean Topic BLEURT	Mean User BLEURT	Distinct-1	Entropy-1	Mean Text Length
gpt-neo-1.3B	**en**	0.17	0.18	0.58	6.72	286.86
	fa	0.06	0.05	0.7	6.54	192.87
	hi	0.1	0.11	0.7	6.45	176.78
	ur	0.07	0.06	0.74	6.4	151.74
	zh	0.13	0.11	0.67	6.22	171.28
gpt-neo-125M	**en**	0.17	0.17	0.55	6.39	261.49
	fa	0.07	0.06	0.73	6.22	148.14
	hi	0.11	0.11	0.68	6.16	159.2
	ur	0.08	0.08	0.75	6.24	144.86
	zh	0.13	0.11	0.63	6.07	176
GT	**en**	0.22	0.27	0.85	4.61	38.86
	fa	0.04	0.17	1	4	16
	hi	0.18	0.17	0.75	5.45	65
	ur	0.12	0.13	0.85	4.79	44.91
	zh	0.12	0.09	0.86	4.92	37

Table 7: GEM Metrics by Language

existing text. We accomplish this with multi-objective few-shot prompting and automated evaluation metrics for multiple aspects of text quality. While it is clear that even in best-case scenarios language models like GPT can generate a percentage of text that does not match the given objectives, we believe that a filtering step including multiple evaluation metrics is a good approach to overcoming this limitation. Future plans include the application of domain-adaptive pre-training and platform business rules to improve relevance to social media genre, continuous/soft prompting instead of discrete/hard prompting to further improve the quality of generated text, real world events prompting similar to reply prompting, and multimedia data generation. Other work in this domain could involve testing other, potentially smaller language models, or assessing the threat posed by adversarial or malicious information campaigns that utilize text generation methods, and how best to detect them.

Acknowledgements

We also thank Graham Horwood for his suggestions on evaluation methods and Mark Williams for initial input for the project. We thank our anonymous reviewers for their valuable feedback. This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) and the Army Research Office (ARO) under Contract No. W911NF-17-C-0095. The content of the information in this document does not necessarily reflect the position or the policy of the U.S. Government, and no official endorsement should be inferred..

References

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow.

Mack Blackburn, Ning Yu, John Berrie, Brian Gordon, David Longfellow, William Tirrell, and Mark Williams. Corpus development for studying online disinformation campaign: a narra-

tive+ stance approach. In Proceedings for the First International Workshop on Social Threats in Online Conversations: Understanding and Management, pages 41–47.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. arXiv preprint arXiv:1803.11175.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. CoRR, abs/1810.04805.

Sebastian Gehrmann, Tosin P. Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Aremu Anuoluwapo, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna Clinciu, Dipanjan Das, Kaustubh D. Dhole, Wanyu Du, Esin Durmus, Ondrej Dusek, Chris C. Emezue, Varun Gangal, Cristina Garbacea, Tatsunori B. Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Rubungo Andre Niyongabo, Salomey Osei, Ankur P. Parikh, Laura Perez-Beltrachini, Niranjan Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. 2021. The gem benchmark: Natural language generation, its evaluation and metrics. ArXiv, abs/2102.01672.

Dazhen Wan Zhihong Shao Pei Ke Jian Guan Yilin Niu Xiaoyan Zhu Huang, Fei and Minlie Huang. 2020. Cotk: An open-source toolkit for fast development and fair evaluation of text generation. arXiv:2002.00583.

Hassan Kane, Muhammed Yusuf Kocyigit, Ali Abdalla, Pelkins Ajanoh, and Mohamed Coulibali. 2020. NUBIA: NeUral based interchangeability assessor for text generation. In Proceedings of the 1st Workshop on Evaluating NLG Evaluation, pages 28–37, Online (Dublin, Ireland). Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. arXiv preprint arXiv:2107.13586.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. CoRR, abs/1907.11692.

Marco Lui and Timothy Baldwin. 2012. Langid.py: An off-the-shelf language identification tool. In Proceedings of the ACL 2012 System Demonstrations, ACL '12, page 25–30, USA. Association for Computational Linguistics.

Goran Murić, Alexey Tregubov, Jim Blythe, Andrés Abeliuk, Divya Choudhary, Kristina Lerman, and Emilio Ferrara. 2020. Massive cross-platform simulations of online social networks. In Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems, pages 895–903.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pages 311–318.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In Proceedings of the Third Conference on Machine Translation: Research Papers, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Noam Shazeer Adam Roberts Katherine Lee Sharan Narang Michael Matena Yanqi Zhou Wei Li Raffel, Colin and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research 21 (2020) 1-67.

Thilina Rajapakse. 2020. Simple transformers. `https://github.com/ThilinaRajapakse/simpletransformers`.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics.

Ananya B Sai, Akash Kumar Mohankumar, and Mitesh M Khapra. 2020. A survey of evaluation metrics used for nlg systems. arXiv preprint arXiv:2008.12009.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. ArXiv, abs/1910.01108.

Yifan Hao Andong Jing Dongxin Liu Shengzhong Liu Hoang Hai Nguyen David M. Nicol Huajie Shao Tianshi Wang Shuochao Yao Yu Zhang Omar Malik Stephen Dipple James Flamino Fred Buchanan Sam Cohen Gyorgy Korniss Boleslaw K. Szymanski Tarek Abdelzaher, Jiawei Han. 2020. Multiscale online media simulation with socialcube. Computational and Mathematical Organization Theory.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. CoRR, abs/1906.08237.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In International Conference on Learning Representations.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, pages 1097–1100.

A Appendix: Authorship Verification Scores by Language

Chart 2 shows the average of the score used for authorship verification on GT positive (real texts by the same user), GT negative (real texts by different users) and synthetic texts, of various languages.

B Reply Validity Scores by Language

Chart 3 shows the average of the score used for reply validity on GT positive (real statement-reply pairs), GT negative (mismatched real texts that are not statement-reply pairs) and synthetic replies, of various languages.

C Statement-Reply Language Counts

Chart 4 shows the count of tweet-reply pairs by language of original tweet and reply. The high count on the diagonal means that tweets in one language are frequently replied to with replies in the same language.

D Synthetic Text Examples on Topic

Synthetic text with intended topic "controversies/china/border" (India-China border disputes) with Topic BLEURT score 0.38:
"The Tension continues between India and Beijing A #ChinaComesModiRuns official tweet says that Beijing is the only government in the world that is trying to bully India into giving way to China's agenda. @Modi ran a tweet saying that China will not be allowed to gain a better foothold in the Indian Ocean. @Modi ran a tweet saying that China will not be allowed to gain a better foothold in the Indian Ocean."

Synthetic text with intended topic "benefits/development/energy" (energy development projects in Pakistan) with Topic BLEURT score 0.27:
"by @Carr-Ebrahim @JaxyIvan 2069 @Manssazh 2069 This project is to provide for power generation facility from PJM. It will provide thermal power system from PJM. The project will be the first industrial power plant in the world to be operated by either one

Model	Language	Cosine Sim GT User Centroid	Std Dev. Cosine Sim	Relevant Text Count
The method used to generate the text	*Language code*	*Average cosine similarity of generated text to the centroid of the user's text in a GT eval set*	*Standard Dev. of the cosine similarities*	*Number of texts used in evaluation*
GT Positive	ta	0.98	0.01	40
GT Positive	de	0.96	0.05	142
GT Positive	es	0.96	0.04	108
GT Positive	bn	0.95	0.07	48
GT Positive	ja	0.95	0.06	132
GT Positive	hi	0.95	0.06	4878
GT Positive	pnb	0.94	0.05	55
EleutherAI/gpt-neo-125M	hi	0.94	0.03	62
GT Positive	fa	0.94	0.06	76
GT Positive	ur	0.93	0.07	4258
GT Positive	en	0.92	0.08	30555
GT Positive	zh	0.90	0.08	373
EleutherAI/gpt-neo-125M	ur	0.89	0.10	121
EleutherAI/gpt-neo-1.3B	en	0.84	0.13	985
EleutherAI/gpt-neo-125M	ta	0.84	0.04	63
EleutherAI/gpt-neo-125M	en	0.83	0.13	1042
EleutherAI/gpt-neo-125M	fa	0.83	0.13	71
EleutherAI/gpt-neo-125M	zh	0.83	0.14	49
GT Negative	es	0.81	0.14	108
EleutherAI/gpt-neo-1.3B	ur	0.81	0.09	69
GT Negative	de	0.80	0.17	142
EleutherAI/gpt-neo-1.3B	zh	0.75	0.18	51
GT Negative	fa	0.70	0.19	76
GT Negative	en	0.67	0.20	30555
GT Negative	pnb	0.66	0.22	55
GT Negative	ta	0.63	0.16	40
GT Negative	zh	0.63	0.18	373
GT Negative	ja	0.62	0.23	132
GT Negative	bn	0.62	0.12	48
GT Negative	ur	0.59	0.22	4258
GT Negative	hi	0.53	0.19	4878

Figure 2: Authorship Verification Scores by Language

Model	Language	NSP Classifier Output Mean	Std Dev NSP Classifier Output	NSP Classifier Prediction Mean	Std Dev NSP Classifier Prediction	Relevant Text Count
Method used to generate the text	*Language code*	*Average of raw output of the NSP classifier for the statement-reply pairs. Higher is better.*	*Standard dev. of the NSP classifier outputs*	*Average of the prediction of the NSP classifier for the statement-reply pairs. Range 0-1. Higher is better*	*Standard dev. of the NSP classifier predictions*	*Number of texts used in the evaluation*
GT_positive	ur	0.37	0.53	0.91	0.28	645
GT_positive	hi	0.34	0.53	0.93	0.26	151
GT_positive	en	0.33	0.54	0.89	0.31	1674
EleutherAI/gpt-neo-1.3B	hi	(0.41)	1.32	0.66	0.47	136
EleutherAI/gpt-neo-125M	hi	(0.64)	1.40	0.56	0.50	48
EleutherAI/gpt-neo-1.3B	kn	(0.99)	1.18	0.31	0.47	65
EleutherAI/gpt-neo-125M	ur	(1.05)	1.32	0.37	0.48	628
EleutherAI/gpt-neo-1.3B	ur	(1.16)	1.29	0.32	0.47	1435
EleutherAI/gpt-neo-125M	en	(1.18)	1.22	0.32	0.47	1193
EleutherAI/gpt-neo-125M	fa	(1.23)	1.22	0.27	0.44	395
EleutherAI/gpt-neo-1.3B	en	(1.26)	1.24	0.31	0.46	973
EleutherAI/gpt-neo-1.3B	fa	(1.26)	1.10	0.24	0.43	142
EleutherAI/gpt-neo-125M	pnb	(1.27)	1.25	0.29	0.45	77
EleutherAI/gpt-neo-1.3B	ta	(1.32)	1.18	0.23	0.42	53
GT_negative	hi	(1.42)	1.56	0.37	0.48	160
EleutherAI/gpt-neo-1.3B	ar	(1.43)	1.10	0.18	0.38	68
EleutherAI/gpt-neo-1.3B	pnb	(1.43)	1.18	0.21	0.41	56
GT_negative	en	(1.50)	1.51	0.32	0.47	1566
GT_negative	ur	(1.60)	1.58	0.34	0.47	718

Figure 3: Reply Validity Scores by Language

Synthetic Reply Language \ Real Tweet Language	ar	en	es	fa	gu	hi	ja	kn	mr	pnb	ru	ta	te	ur	zh
ar	1	83	0	0	0	2	0	0	1	0	0	0	0	3	2
en	1	2025	1	0	0	137	0	0	4	1	1	1	1	210	4
es	0	23	0	0	0	1	0	0	0	0	0	0	0	3	0
fa	1	410	1	2	0	18	0	0	0	3	0	1	0	101	1
gu	0	25	0	0	2	5	0	0	0	0	0	0	0	0	0
hi	0	89	0	0	0	115	0	0	4	0	0	0	0	7	1
ja	0	19	0	0	0	1	1	0	0	0	0	0	0	3	0
kn	0	55	0	0	0	5	0	7	2	0	0	0	0	3	0
mr	0	7	0	0	0	1	0	0	3	0	0	0	0	0	0
pnb	0	108	0	0	0	2	0	0	0	2	0	0	0	32	0
ru	0	32	0	0	0	3	0	0	0	0	0	0	0	3	0
ta	0	61	0	0	0	13	0	0	0	0	0	12	0	4	0
te	0	23	0	0	0	1	0	0	0	0	0	1	3	0	1
ur	1	1455	2	3	0	65	1	0	1	6	0	0	0	719	0
zh	0	40	0	0	0	3	0	0	0	0	0	0	0	1	4

Figure 4: Statement-Reply Language Counts

company. #Pakistan#Polarity #UK-Indian Union of Petroleum, Air and Chem"

E Synthetic Text Examples in a User Style

Real user bio:
"News. Views. Analysis. Conversations. India's No.1 digital news destination, world's largest-selling English newspaper - A Times Internet Limited Product"
Real tweet from user:
"Pakistan PM Imran Khan claims PM Modi apologised for lockdown, Pak media corrects him url: https://t.co/m4vAITPOTGhdHATzA-KxaIg "

Synthetic tweets:
1. *"India's Prime Minister Narendra Modi apologises to Pakistan PM Imran Khan after a video clip of him being filmed during the crisis. The video has gained support from international organisations url: https://t..."*
2. *"The Government of Pakistan has made efforts, I am pleased to say, to ensure that the Indian government did not see the rise of Pakistans internal tensions," Foreign Minister Shah Mehmood Qureshi tells the New York Times. "The two nations have a very good relationship." He also claims Pakistan has never asked for India's "aid and support" Read: https://t.co/eMhXC7Lr1mDnT3fJL"*