Measuring the Effect of Influential Messages on Varying Personas

Anonymous ACL submission

Abstract

Predicting how a user responds to news events enables important applications such as allowing intelligent agents or content producers to estimate the effect on different communities and revise unreleased messages to prevent unexpected bad outcomes such as social conflict and moral injury. We present a new task, Response Forecasting on Personas for News Media, to estimate the response a persona (characterizing an individual or a group) might have upon seeing a news message. Compared to the previous efforts which only predict generic comments to news, our task not only introduces personalization in the modeling but also predicts sentiment polarity and intensity of each response, enabling more accurate and comprehensive inference on the mental state of the persona. At the same time, the sentiment dimensions made evaluation more reliable. We create the first dataset, which consists of 13,357 responses to 3,847 news headlines from Twitter, for evaluating this new task. We further evaluate the SOTA neural language models with our dataset, and the empirical results suggest that the included persona attributes are helpful for the performance of all response dimensions. Our analysis shows that the best-performing models are capable of predicting responses that are consistent with the personas, and as a byproduct, the task formulation also enables many interesting applications in the analysis of social network groups and their opinions, such as the discovery of extreme opinion groups. We will release the dataset, repository, and models for research purposes.

1 Introduction

To prevent the flooding of misinformation and hate speech on the internet, a great amount of progress has been made toward identifying and filtering such content on social media using machine learning models (Fung et al., 2021; Su et al., 2022; ElSherief et al., 2021; Sap et al., 2019). While directly

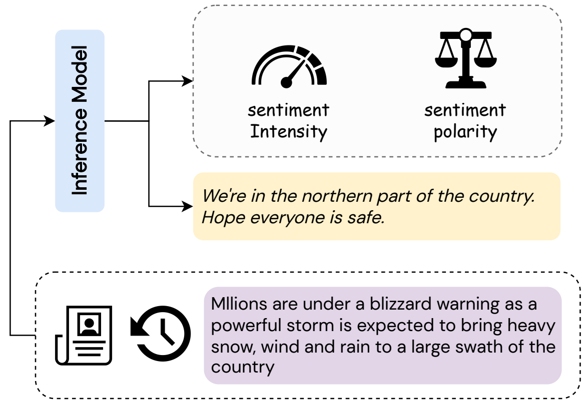

Figure 1: An example illustrating the task. The input consists of persona attributes (e.g., historical activities and profile) and a news message. The model is asked to predict response in multiple dimensions.

creating message-level labels is a natural way to address the issue, it is equally important to measure the influence of the message on different viewers as a way to decide how to manage the publication of the messages.

Existing efforts (Lin and Chen, 2008; Giachanou et al., 2018; Yang et al., 2019) have made steps toward predicting population-level response (e.g., what is the most likely response to a news message) for news, but neglected the importance of personas in measuring influence. According to Individual Differences Theory (Riley, 1959), which proposes that individuals respond differently to the mass media according to their psychological needs, the same message can impact different population groups/personas in different ways. For example, a message claiming the honor of sacrificing others' lives for a religious goal might still agitate people who are prone to agree to such messages. It is therefore essential to consider personalization when inferring viewers' responses.

On the otherhand, the previous approaches that predict text-level responses (Yang et al., 2019; Wu et al., 2021) have only used generation metrics for

Split	Train	Dev.	Test
# Samples	10,977	1,341	1,039
# Headlines	3,561	1,065	843
# Users	7,243	1,206	961
Avg # Profile Tokens	10.75	11.02	10.50
Avg # Response Tokens	12.33	12.2	11.87
Avg # Headline Tokens	19.79	19.82	19.72

Table 1: Summary statistics for the dataset.

automatic evaluation, yet the same sentiment can be expressed in a multitude of ways, and text alignment metrics like BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) do not credit cases where the sentiments match but semantics do not align well. As a result, it is crucial to evaluate the sentiment dimensions of user responses.

We propose Response Forecasting on Personas for News Media, a task for measuring the influence of news media messages on viewers by predicting viewers' responses. In particular, the input consists of the news message and persona information (e.g., user profile and history in our dataset), and we define response in terms of sentiment polarity, sentiment intensity, and textual response. While we include three categories in this work, many other interesting aspects can also be defined (e.g., change of altitude toward real-world entities) and we leave them to future work. Studying the problem of forecasting individual viewers' responses allows the creation of tools to assist analysts and online content producers to estimate the potential impact of messages on different communities, and sheds light on new applications such as automatically re-writing a message/email to achieve a communication goal (e.g., to obtain a positive response from the receiver). Furthermore, this new task also helps to understand associations between user attributes and emotional responses.

To construct a test bed for this task, we collect a dataset from Twitter consisting of 13,357 labeled responses to 3,847 news headlines from Twitter. Using the corpus, we examine how state-of-the-art neural models work in our task. We find that the models can predict responses with reasonable accuracy yet still have a large room for improvement. We also find that the best-performing models are capable of predicting responses that are consistent with the personas, indicating that the models may be used for many exciting applications such as the discovery of groups with different opinions.

2 Dataset Collection

In this section, we describe how we construct data from Twitter. Specifically, we used Twitter API[1] to crawl news headlines and comments below each headline from CNN Breaking News[2], which is one of the most popular news accounts on Twitter.

Preprocess. We collected news headlines and corresponding comments from CNN Breaking News between January 2017 and January 2019 and removed the comments that are over 50 tokens to avoid spamming. We stripped away HTML syntax tokens and normalized user reference with special tokens "@user".

2.1 Persona Data

We refer to the users who produce the comments as responders. To describe responders, we collected different classes of persona attributes from Twitter. Specifically, we collected (1) User Profile, which is a short paragraph describing the user, and (2) User History, which are tweets written directly by the user. For user historical posts, to ensure that future posting activities are not included when predicting the comment, we collect the historical posts prior to the earliest data sample in our dataset for each individual user.

2.2 Annotation

We obtained 14k headline and comment pairs from preprocessing. In the annotation stage, we collect labels for sentiment intensity and polarity of comments based on the context of the headline. For the 10k training instances, we produce automatic labels using deep-learning models trained on existing message-level datasets. More specifically, we train a Deberta-based model (He et al., 2020) using data from SemEval-2018 Task 1[3] (Mohammad et al., 2018), reaching over 85% Pearson correlation. We then proceed to use crowd-sourcing to annotate the remaining 2k samples as our evaluation set.

Task Setup. The annotation for the evaluation set is performed using the Amazon Mechanical Turk (MTurk) crowd-sourcing platform. The workers were each asked to annotate a headline and comment pair with three workers assigned to each data sample. During the annotation, the annotator is asked to select the sentiment polarity label and

[1] developer.twitter.com/en/docs/twitter-api
[2] twitter.com/cnnbrk
[3] https://competitions.codalab.org/competitions/17751

Name	BLEU	BScore	Meteor	R-1	R-L	Avg. Len	ϕ_{int} r_s	ϕ_{int} r	ϕ_p MiF1	ϕ_p MaF1
Majority	-	-	-	-	-	-	-	-	43.41	20.18
Random	-	-	-	-	-	-	0.62	0.41	35.51	30.55
GPT2	1.59	-5.78	3.36	6.50	1.90	9.64	50.34	49.78	60.25	56.85
T5	6.95	-5.71	5.98	**10.40**	**2.70**	18.87	50.06	49.26	63.72	57.85
BART	**8.17**	**-5.67**	**6.09**	9.90	2.50	**21.05**	**62.03**	**61.82**	**67.85**	**63.23**
BART w/o Profile	7.30	-5.70	5.91	10.00	2.50	19.47	57.95	58.20	67.28	62.26
BART w/o History	5.24	-5.88	4.41	7.70	1.50	18.62	48.80	48.63	59.00	53.29
BART w/o Both	3.90	-5.92	4.00	7.90	1.80	15.73	45.28	44.75	61.41	46.01

Table 2: Response forecasting results above show that the state-of-the-art models can predict responses with reasonable performance. The best overall performance is bolded.

the intensity of the sentiment based on their understanding of the input. The workers select positive, negative, or neutral for the sentiment polarity label and select on the integer scale of 0 to 3 for intensity. Four hundred and fifteen workers participated in this task in total and all annotators are paid a fair wage above the federal minimum.

Quality Control. To ensure the quality of annotation, we allowed only the workers who have at least 95% approval rate and have had at least 5,000 hits approved to access our tasks. We further removed workers who have a <70% accuracy in the first 30 annotations and discarded the assignments that have completion time deviated from the expected average largely. We used majority voting to determine the final labels: if at least two annotators agreed on a label, we chose it as the final label. The resulting annotated samples achieve an inter-annotator agreement accuracy of 81.3%. We show the statistics of the dataset in Table 1.

3 Response Forecasting on Personas for News Media

3.1 Task Formulation

In this task, we aim to predict sentiment polarity, sentiment intensity, and textual response from an individual when the individual sees a message on news media. Formally, given persona $\mathcal{P}$ (represented by profile, or historical posts), and a source message $\mathcal{M}$, the task is to predict the persona's sentiment polarity ϕ_p (i.e., *Positive*, *Negative*, *Neutral*) and sentiment intensity ϕ_{int} (i.e., in the scale of 1 to 3), and textual expression t. Our goal is to encode $\mathcal{P}$ and produce ϕ_p, ϕ_{int}, and t at decoding time. We formulate the task as a conditional generation problem and use the following maximum-likelihood objective to train a generative model:

$$\sum_{i}^{N} \log p(O_i|O_{<i-1}, \mathcal{P})$$

where O is the output string concatenating ϕ_p, ϕ_{int}, and t with special separator tokens.

Model	Persona	Label	Context
GPT2	3.18	3.84	2.84
T5	3.68	4.23	3.57
BART	**4.35**	**4.42**	**3.99**

Table 3: The table shows human evaluation results based on three consistency measures, supporting the automatic evaluation findings.

3.2 Experimental Setup

For deep learning-based text generators, we fine-tune decoder-only text generator GPT2 (Radford et al., 2019) as well as two Encoder-Decoder models T5 (Raffel et al., 2019) and BART (Lewis et al., 2019). Greedy decoding is used for all the models during training. We further perform ablation on the best-performing model by removing different user attributes. We further include two naive baselines, *Random* and *Majority*, for sentiment dimensions, where each prediction follows either the majority label or a random label. Our neural models are implemented using Pytorch (Paszke et al., 2019) and Huggingface Transformers (Wolf et al., 2020). The reproducibility and hyperparameter details can be found in Appendix Table 4.

3.2.1 Evaluation Metrics

Automatic. We use BARTScore (Yuan et al., 2021), BLEU (Papineni et al., 2002) , METEOR (Banerjee and Lavie, 2005), and ROUGE (Lin, 2004) to evaluate textual response generation performance. Note that BARTScore computes the log-likelihood of producing the reference text given the generated text using a BART model pretrained on ParaBank2[4]. Furthermore, we use Pearson and Spearman correlation to evaluate sentiment intensity, and F1 to evaluate sentiment polarity.

Manual. We conduct human evaluation to measure the consistency of the generated outputs from those models. We define three types of consistency metrics: (1) *persona consistency*: whether the output reflects the persona's characteristics, (2) *label consistency*: whether the response text and sentiment are consistent with each other, (3) and *context consistency*: whether the output is responding to the input news headline. We randomly select 10 personas with distinct characteristics (i.e., the writing style/interest/profession do not clearly overlap) and 10 news headlines from distinct topics, and consequently generate 100 responses using each model. The samples are distributed to 5 raters who score each output based on our metrics. The raters are master students who passed a small quiz of 20 samples with at least 80% accuracy. We additionally make sure that each rater is familiar with the persona information (e.g., profile and history) before starting to work on the task.

3.3 Results

Automatic Evaluation. Across the metrics in Table 2, we can see that BART provides us with the highest quality response predictions on both sentiment and text levels. As expected, the performance of simple baselines is relatively low compared to other models, showing that the dataset does not have a class imbalance issue. While the automatic generation scores are generally low (i.e., words do not align well), the sentiment prediction scores are much higher in scale, demonstrating the importance of sentiment scoring to make a fair judgment of the result; the model needs to be credited for correctly predicting the latent sentiment even if it does not utter the exact sentence. Finally, we ablate user attribute features one by one. As shown in the table, not only both features included are effective for the task, but they are also complementary of each other.

Human Evaluation. The results from human judgments (Table 3) in general support the automatic evaluation findings. Among all three models, our approach with BART reaches the highest on all metrics, showing it can generate responses of better quality than others. The difference between models on Label Consistency is noticeably lower than other metrics, and the number suggests that pretrained language models are capable of producing sentiment labels consistent with the textual expression. On the other hand, we find that BART can produce responses more consistent with the controllable variables than GPT2, which might be attributed to its denoising pretraining (e.g., it adapts better to different modeling formats). In fact, the outputs show that GPT2 hallucinates more often than other models.

3.4 Application

We hypothesize that the formulation of the task enables the application of discovering groups with different opinions on issues. We verify the hypothesis by collecting personas with contrasting stances on an issue and generating responses based on this issue. We find that the output from the model stays consistent with the persona (examples are shown in the Appendix Table 5). The result demonstrates the potential for application on social network analysis. Since the model is able to generalize to different personas or news, an analyst can therefore replace the news headline with others to segment the population based on different issues, or manually construct a persona to visualize how a person from a particular community would respond to certain issues.

4 Conclusions and Future Work

We propose Response Forecasting on Personas for News Media, a new task that tests the model's capability of estimating the responses from different personas. The task enables important applications such as estimating the effect of unreleased messages on different communities as an additional layer of defense against unsafe information (e.g., information that might cause conflict or moral injury). We also create the first dataset for evaluating this new task and present an evaluation of the state-of-the-art neural models. The empirical results show that the best-performing models are able to predict responses with reasonable accuracy and produce outputs that are consistent with the personas. The analysis shows that the models are also able to generate contrasting opinions when conditioned on contrasting personas, demonstrating the feasibility of applying the models to discovering social groups with different opinions on issues for future work.

[4]https://github.com/neulab/BARTScore

Limitations

While the training method makes use of user profile description and history, one additional factor that is important is the structure between users and news articles. Knowing a user's social circles can often give hints about the user's interests and beliefs, which can potentially help the model to infer how a particular persona would respond to an issue. A possible direction is to design a method that explores the social context features (e.g., social network) via graph-based algorithms.

Ethics

During annotation, each worker was paid $15 per hour (converted to per assignment cost on MTurk). If workers emailed us with any concerns, we responded to them within 1 hour. The research study has also been approved by the Institutional Review Board (IRB) and Ethics Review Board at the researchers' institution. Regarding privacy concerns our dataset may bring about, we follow the Twitter API's Terms of Use[5] and only redistribute content for non-commercial academic research only. We will additionally require interested parties to sign an agreement on data and model usage to make sure the resource will be used in ethical ways and available for academic use only.

References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech.

Yi Fung, Christopher Thomas, Revanth Gangi Reddy, Sandeep Polisetty, Heng Ji, Shih-Fu Chang, Kathleen McKeown, Mohit Bansal, and Avirup Sil. 2021. Infosurgeon: Cross-media fine-grained information consistency checking for fake news detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1683–1698.

[5] https://developer.twitter.com/en/developer-terms/agreement-and-policy

Anastasia Giachanou, Paolo Rosso, Ida Mele, and Fabio Crestani. 2018. Emotional influence prediction of news posts. In *Twelfth International AAAI Conference on Web and Social Media*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Kevin Hsin-Yih Lin and Hsin-Hsi Chen. 2008. Ranking reader emotions using pairwise loss minimization and emotional distribution regression. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 136–144, Honolulu, Hawaii. Association for Computational Linguistics.

Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. 2019. On the variance of the adaptive learning rate and beyond. *arXiv preprint arXiv:1908.03265*.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 1–17.

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.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

John W Riley. 1959. *Mass communication and the social system.*

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Ting Su, Craig Macdonald, and Iadh Ounis. 2022. Leveraging users' social network embeddings for fake news detection on twitter.

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

Yuwei Wu, Xuezhe Ma, and Diyi Yang. 2021. Personalized response generation via generative split memory network. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1956–1970, Online. Association for Computational Linguistics.

Ze Yang, Can Xu, Wei Wu, and Zhoujun Li. 2019. Read, attend and comment: A deep architecture for automatic news comment generation.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.

A Appendix

A.1 Implementation Details

We implement the models using the 4.8.2 version of Huggingface Transformer library[6](Wolf et al., 2020). We use Oct 1, 2021 commit version of the BART-base model (139M parameters) from Huggingface[7]. We use Huggingface datasets[8] for automatic evaluation metrics. The BART Score comes from the author's repository[9] and we used the one trained on ParaBank2. The hyperparameters for the experiment are shown in Table 4 (applied to all models) and the ones not listed in the table are set to be default values from the transformer library. We use RAdam (Liu et al., 2019) as the optimizer. We perform hyperparameter search on the batch size from {16, 32}, pretrained language model learning rate from {3e-5, 4e-5, 5e-5}. We perform our experiments on 32 GB V100. The experiments can take up to 15 hours.

Name	Value
seed	42
learning rate	5e-5
batch size	16
weight decay	5e-4
RAdam epsilon	1e-8
RAdam betas	(0.9, 0.999)
scheduler	linear
warmup ratio (for scheduler)	0.06
number of epochs	20
metric for early stop	SacreBLEU[10]
patience (for early stop)	15
length penalty	1.2
beam search size during eval	5

Table 4: Hyperparameters. The ones below the mid-line are generation related.

[6] https://github.com/huggingface/transformers

[7] https://huggingface.co/facebook/bart-base/commit/ea0107eec489da9597e9eefd095eb691fcc7b4f9

[8] https://github.com/huggingface/datasets

[9] https://github.com/neulab/BARTScore

Headline: Millions are under a blizzard warning as a powerful storm is expected to bring heavy snow, wind and rain to a large swath of the country	
Purity & Love	Degradation
We're in the northern part of the country. Hope everyone is safe	Mother Nature sure is pissed off at us

Headline: Judge says Trump may have been urging supporters to 'do something more' than protest on Jan. 6	
Pro-President Trump	Anti-President Trump
The liberal media & Dems are always negative when it comes to anything. They don't care about anything except themselves	Hahahahahahaha! They figured that Trump would be impeachedby now! But the traitorous Republicans are slowing down the process.

Headline: Russia and Ukraine are at war	
Pro-Russia	Pro-Ukraine
Support Russia	Support Ukraine

Table 5: Tables showing different cases that contrasting the persona (selected from existing) can lead to the generation of contrasting opinions on issues. For each table, the middle row contains different personas, and the third row contains the responses from each persona.