Hate2Explain: Crowdsourced Explanations as a Cultural Bridge in Understanding Hateful Memes

Chao-Chun Han[1], Yi-Ching (Janet) Huang[2], Nanyi Bi[1], Jane Yung-jen Hsu[1]
[1] National Taiwan University
[2] Eindhoven University of Technology
r08922127@csie.ntu.edu.tw, y.c.huang@tue.nl, nanyibi@ntu.edu.tw, yjhsu@csie.ntu.edu.tw

Abstract

Detection of hateful memes depends on semantic understanding of the juxtaposition of short texts over image(s). Independently innocent texts or images may become hateful when they are combined in specific ways, which can be tricky for people without knowledge of the cultural context to understand. This work presents a new approach to generating explanations that help bridge the cultural gap in understanding hateful memes. Inspired by prior research, a three-stage crowdsourcing workflow is proposed to guide crowd workers to generate, annotate, and revise explanations of hateful memes. To ensure the quality of explanations, a self-assessment rubric is designed to evaluate the explanations using four criteria: target, clarity, explicitness, and utility. We evaluated the proposed workflow in an online study with 66 participants, compared to a single-stage workflow. The results showed that the three-stage workflow guided crowds to generate explanations that meet more criteria than the explanations generated by a single-stage workflow.

Introduction

The rapid spread of hateful memes has caused serious social problems, such as exacerbating ethnic tensions (Mathew et al. 2019; Gelber and McNamara 2016). While great efforts have been made in developing various approaches to detecting hateful memes automatically (Zhou, Chen, and Yang 2021; Zhu 2020), we still face tremendous challenges when dealing with such diverse forms of hate speech. There is no uniform definition of hate speech (Burnap and Williams 2015; Baider, Assimakopoulos, and Millar 2017) and it might be affected by context, semantics, and social factors. People's perceptions are also different (Salminen et al. 2018, 2019). To address these challenges, this work aims to explore a new approach to improved understanding and awareness of hateful memes.

Figure 1 shows the proposed three-stage workflow for generating explanations with crowd workers. First, we design the Generate-Annotate-Revise workflow, a multi-stage workflow that breaks a complex explanation task into a series of micro-tasks. Second, a self-assessment rubric is designed based on the common traits of hate speech. The aim of the rubric is to allow crowd workers to self-assess whether their written explanations meet the criteria in the Generate and Revise stages.

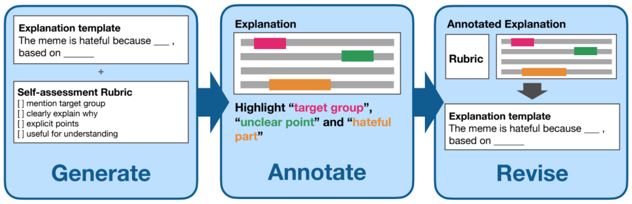

Figure 1: The Generate-Annotate-Revise workflow.

To evaluate our approach, we deployed the workflow and conducted an online study on Amazon Mechanical Turk. The goal of this study is to evaluate the quality of explanation generated by our workflow compared to a single-stage workflow. The results showed that our workflow generated explanations that satisfy more criteria than the explanations generated by a single-stage workflow.

Related Work Recent work has shown that the state-of-the-art multi-modal hateful meme detections (Lu et al. 2019; Li et al. 2019) performed worse than humans (69.47% vs. 84.7% accuracy) because the machine learning algorithms have difficulty identifying hate speech due to its protean forms. Even human users need to have adequate prior knowledge to determine whether the content is hateful or not. Laaksonen et al. also showed that labeling hate speech with a consensus is rather difficult (Laaksonen et al. 2020). Hence, there is a need to have comprehensible explanations to reveal the reason and hidden background information to help people understand hateful memes.

Prior work has used crowdsourcing workflow to generate explanations for humorous memes. Lin et al. applied the linguistic theory into workflow design and created useful explanations for humorous memes (Lin, Huang, and Hsu 2014). Differing from prior work, this research aims to generate explanations for specific types of memes—hateful memes. Due to the high diversity of hateful memes, we need a more scalable approach to collect explanations. Inspired by prior crowdsourcing research (Bernstein et al. 2010; Dow et al. 2012), we proposed a multi-stage workflow with self-assessment criteria; the criteria are designed based on the common traits of hate speech extracted from prior literature (Sellars 2016; Fortuna and Nunes 2018).

Workflow Design

Assessment Rubric Hate speech does not have a universal definition (Burnap and Williams 2015; Baider, Assimakopoulos, and Millar 2017), but most researchers agree that two main traits are critical: (1) the target group and (2) the hate content (Sellars 2016; Fortuna and Nunes 2018). The target group is the primary threshold factor to separate "hate speech" from any other form of harmful speech (e.g., bullying or threats). Hate speech usually targets a group or an individual related to a group. The content containing hatred is the key to identifying whether the speech is hateful or not. Also, people intentionally misspell words or use pejorative or metaphors to express hatred (Baider, Assimakopoulos, and Millar 2017). Such implicit and complicated expressions make it difficult to determine if there is an element of hatred. To support crowd workers to assess the hate speech, we used the two common traits of hate speech to design an assessment rubric. In total, we incorporate four criteria in the assessment rubric, including target, clarity, explicitness, and utility. Target is used to check whether an explanation refers to a target group. Clarity and explicitness are designed to assess whether an explanation clearly indicates the cause of hatred and explicitly expresses implicit meaning. Utility is used to evaluate the overall usefulness of the explanation.

Generate-Annotate-Revise Workflow Our proposed multi-stage workflow consists of three stages: 1) Generate, 2) Annotate, and 3) Revise.

Stage 1: Generate: The goal of the Generate stage is to collect original explanations. The task asks workers to read one hateful meme and generate explanations according to a given template. The template has two parts: explanation and background information. The first step is the explanation step ("The meme is hateful because"), the aim of which is to have crowd workers explain why they think the meme is hateful. The second step is the background step ("Based on the information below"). The crowd workers will write down the background information that justifies their judgment. This information could be in or out of the meme itself and therefore helps contextualize the memes by giving background information. In the end, crowd workers will self-assess their explanations based on the assessment rubric and revise the explanations to meet the four criteria.

Stage 2: Annotate: The goal of the Annotate stage is to examine whether explanations satisfy the assessment criteria by annotating three components in the given explanations. In this stage, the crowd workers are asked to annotate the target groups, unclear areas, and hateful parts by highlighting words in the explanation. These three different annotations are indicated by different background colors. In the end, the workflow integrates all annotations from three distinct crowd workers and generates an annotated explanation.

Stage 3: Revise: The goal of the Revise stage is to allow crowd workers to revise explanations generated by prior workers toward better quality. This task shows one annotated explanation from the previous stage. The annotated explanation indicates the target groups in yellow and the points of the hatred in green. The unclear area in of explanation is marked with a red squiggly line. The crowd workers are asked to modify the annotated explanation based on the assessment rubric.

Preliminary Experiment

We conducted an online study to evaluate the effectiveness of our proposed multi-stage workflow compared to a single-stage workflow (baseline). For the single-stage workflow, we only used the explanation template without a self-assessment rubric to generate explanations.

We selected 15 memes from Hateful Meme Dataset (Kiela et al. 2020) and deployed the two workflows on Amazon Mechanical Turk (MTurk) to generate two types of explanations. In total, 59 crowd workers generated 61 explanations by the multi-stage workflow, and 17 crowd workers generated 60 explanations by the single-stage workflow. Each worker earned $0.15 for generation, $0.10 for annotation, and $0.20 for revision in the multi-stage workflow; they earned $0.15 for the generation task in the single-stage workflow. Eventually, we got 121 explanations.

Results We evaluated these 121 explanations with 66 participants on MTurk. Each explanation was rated by three people on a scale of 1 (poor) to 5 (excellent) for the overall quality as well as the four criteria in the assessment rubric. Each participant earned $0.06 for this evaluation task. Then, we calculated the inter-coder agreement on the evaluation results to filter the erratic evaluations. Finally, we selected the explanation with a fair agreement score with three workers (Mean Cohen's kappa > 0.3), resulting in 25 explanations for multi-stage workflow and 17 explanations for single-stage workflow. We used the Kruskal–Wallis test to analyze the data for two types of crowd explanations. The results showed that there was no significant difference between the multi-stage workflow and the single-stage workflow on the overall rating. However, the explanations generated by our workflow (M = 2.96, SD = 0.6) meet more criteria than the explanations generated by the single-stage workflow (M = 2.19, SD = 1.05), $\chi^2 = 5.92, p < 0.05$.

Discussion and Conclusion

This paper presents the Generate-Annotate-Revise workflow to provide explanations for hateful memes. The assessment rubric, derived from the common traits of hate speech, is used to guide crowd workers to assess the explanation with four criteria. Results from the preliminary study showed that the multi-stage workflow generated explanations that meet more criteria than explanations generated by the single-stage workflow. However, it is important to know that the current results are based on crowd assessments. The high variance of crowd assessments might be an issue to evaluate the quality of explanation. Also, we excluded some memes from this study due to disagreements from multiple workers. Therefore, our next step is to incorporate experts' opinions to evaluate the explanations generated by our workflow. To push this work forward, we plan to conduct a series of studies to investigate how these explanations affect people's perceptions and behaviors.

References

Baider, F. H.; Assimakopoulos, S.; and Millar, S. L. 2017. Hate speech in the EU and the CONTACT project. *Online Hate Speech in the European Union: A Discourse-Analytic Perspective, eds S. Assimakopoulos, FH Baider, and S. Millar (Cham: Springer)* 1–6.

Bernstein, M. S.; Little, G.; Miller, R. C.; Hartmann, B.; Ackerman, M. S.; Karger, D. R.; Crowell, D.; and Panovich, K. 2010. Soylent: a word processor with a crowd inside. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, 313–322.

Burnap, P.; and Williams, M. L. 2015. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & internet* 7(2): 223–242.

Dow, S.; Kulkarni, A.; Klemmer, S.; and Hartmann, B. 2012. Shepherding the crowd yields better work. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*, 1013–1022.

Fortuna, P.; and Nunes, S. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)* 51(4): 1–30.

Gelber, K.; and McNamara, L. 2016. Evidencing the harms of hate speech. *Social Identities* 22(3): 324–341.

Kiela, D.; Firooz, H.; Mohan, A.; Goswami, V.; Singh, A.; Ringshia, P.; and Testuggine, D. 2020. The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. In *Advances in Neural Information Processing Systems*, volume 33, 2611–2624.

Laaksonen, S.-M.; Haapoja, J.; Kinnunen, T.; Nelimarkka, M.; and Pöyhtäri, R. 2020. The Datafication of Hate: Expectations and Challenges in Automated Hate Speech Monitoring. *Frontiers in Big Data* 3.

Li, L. H.; Yatskar, M.; Yin, D.; Hsieh, C.; and Chang, K.-W. 2019. VisualBERT: A Simple and Performant Baseline for Vision and Language. *arXiv preprint* arXiv:1908.03557.

Lin, C.-C.; Huang, Y.-C.; and Hsu, J. Y.-j. 2014. Crowdsourced Explanations for Humorous Internet Memes Based on Linguistic Theories. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 2(1): 143–150.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, 13–23.

Mathew, B.; Dutt, R.; Goyal, P.; and Mukherjee, A. 2019. Spread of hate speech in online social media. In *Proceedings of the 10th ACM conference on web science*, 173–182.

Salminen, J.; Almerekhi, H.; Kamel, A. M.; Jung, S.-g.; and Jansen, B. J. 2019. Online hate ratings vary by extremes: A statistical analysis. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, 213–217.

Salminen, J.; Veronesi, F.; Almerekhi, H.; Jung, S.-G.; and Jansen, B. J. 2018. Online hate interpretation varies by country, but more by individual: A statistical analysis using crowdsourced ratings. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, 88–94.

Sellars, A. 2016. Defining hate speech. *Berkman Klein Center Research Publication* (20): 16–48.

Zhou, Y.; Chen, Z.; and Yang, H. 2021. Multimodal Learning For Hateful Memes Detection. In *2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, 1–6.

Zhu, R. 2020. Enhance Multimodal Transformer With External Label And In-Domain Pretrain: Hateful Meme Challenge Winning Solution. *arXiv preprint arXiv:2012.08290* .