Using ChatGPT to Enhance Students' Behavior in Social Media via the Moral Foundation Theory

Daniele Schicchi[1,*,†], Apoorva Upadhyaya[2,*,†], Marco Fisichella[2] and Davide Taibi[1]

[1] *Institute for Education Technology, National Research Council of Italy, Palermo, Italy*

[2] *L3S Research Center, Leibniz University, Hannover, Germany*

Abstract

Social media poses numerous dangers, such as the spread of toxic content, hate speech, false information, and moral outrage. In particular, the propagation of moral outrage on social media can result in harmful outcomes like promoting conspiracy theories, violent protests, and political polarization. Teenagers frequently use social media and are especially susceptible to these threats, being exposed to several risks that can impact their lives. Recent studies have confirmed that a combination of human and AI efforts can be highly effective in several domains. In this study, we examine human-AI collaboration's potential and efficacy in detecting morality in online posts. To this end, we conducted a pilot study with teenagers to determine their ability to recognize moral content in online posts. Consequently, we exploited Prompt Engineering to program ChatGPT to recognize morality and comprehend its potential as an intelligent tutor in supporting students in such a context.

Keywords

Human-AI collaboration, social media platforms, moral classification, ChatGPT

1. Introduction

Social media (SM) are not immune to harmful elements such as online bullying, moral outrage, cyber harassment, and hate speech. These toxic components can lead to severe implications in the real world, such as depression and anxiety, especially for young individuals who are more susceptible to the effects of such phenomena [1]. Prior research has placed considerable emphasis on moral sentiments when analyzing social media discourse providing insightful perspectives on significant occurrences such as the practice of social distancing and the spread of misinformation. Evaluating moral sentiments on SM has been proven useful for several tasks such as describing the comparative/cooperative attitude towards the pandemic [2], comprehending public opinions expressed on various societal issues [3], and measuring the social distance between people [4]. Therefore, the influence of moral values cannot be underestimated as they profoundly affect

AIxEDU: 1st International Workshop on High-performance Artificial Intelligence Systems in Education, November 06–09, 2023, Rome, Italy

[*] Corresponding author.

[†] These authors contributed equally.

daniele.schicchi@itd.cnr.it (D. Schicchi); upadhyaya@l3s.de (A. Upadhyaya); mfisichella@l3s.de (M. Fisichella); davide.taibi@itd.cnr.it (D. Taibi)

https://orcid.org/0000-0003-0154-2736 (D. Schicchi); https://orcid.org/0000-0002-5796-8704 (A. Upadhyaya); https://orcid.org/0000-0002-6894-1101 (M. Fisichella); https://orcid.org/0000-0003-0154-2736 (D. Taibi)

CEUR Workshop Proceedings (CEUR-WS.org)

people's outlooks and actions. Whether acknowledged or not, these values significantly shape people's positions and decisions on various issues. As such, it is imperative to scrutinize the prevalent moral sentiments on social media platforms. This measure is essential in safeguarding young individuals from experiences that may negatively impact their well-being [5].

Recent advances in Artificial Intelligence have paved the way for new methodologies to help in education [6, 7]. Recent intelligent systems rely on Large Language Models (LLMs). These systems can provide answers to a wide array of questions across various subject areas. They are designed to be user-friendly, with natural language interfaces that allow users to interact with the system conveniently. ChatGPT [8] is an exceptional chatbot that harnesses the power of LLMs to generate digital content that is both relevant and accurate with impressive speed. Moreover, it ensures a seamless and efficient communication experience by offering text-based responses to user queries, setting it apart as a truly advanced and valuable tool for every type of user.

ChatGPT can be a valuable tool in education. It can help students understand complex concepts, gather preliminary information on research topics, and support writing tasks by generating outlines and improving sentence structure. Since its high performance, ChatGPT is a perfect candidate to enhance the learning experience, and its usage is getting more frequent. The main interface to interact with ChatGPT is a prompt, a set of instructions to program the LLM customizing, enhancing and/or refining its capabilities [9]. A prompt has a significant impact on the behavior of the LLM. It defines the rules and guidelines the system must follow and sets the context for the conversation, indicating to the LLM what information is crucial and what the desired output form and content should be.

In this paper, we aim to analyze the capability of ChatGPT to act as a companion to detect moral content on social media. To achieve this objective, we analyzed how ChatGPT classifies the content when different prompt patterns are used. Using the Moral Foundation Twitter Corpus as the ground truth of our experiment, we evaluated the F1 score to ponder the effects of precision and recall and determine ChatGPT performance in this task.

The paper is organized as follows: in section 2 we give an overview about prompt engineering and the usage of ChatGPT in education. In section 3 we describe the method that have conducted to the results that are analyzed in section 4. Finally, in section 6 we give the final discussion and conclusion.

2. Literature Review

2.1. Importance for Adolescents in Detecting Moral Content on Social Media

Previous literature has applied AI techniques to combat societal issues such as the ones relative to climate change and communal riots resulting from misinformation and hate speech on social media [10, 11, 12, 13]. The presence of immoral and toxic content in online communication eventually lead to harmful effects, especially on teenagers such as teen depression, teen anxiety disorder, mental illness and more[1]. Moreover, a recent survey[2] revealed how social media users

[1] https://www.newportacademy.com/resources/well-being/effect-of-social-media-on-teenagers/

[2] https://www.medienanstalt-nrw.de/fileadmin/user_upload/NeueWebsite_0120/Themen/Hass/Ergebnisbericht_forsa-Befragung_zu_Hate_Speech_im_Internet_2022.pdf

aged 14-24 engage with hate speech due to its entertainment value. Therefore, it is crucial for teenagers to recognize moral/non-moral content in social media and be warned to exercise caution when consuming such content. This prompted us to use the moral task and conduct a pilot study that analyzes teenagers' and AI performance in detecting moral sentiments.

2.2. ChatGPT: Prompt Engineering and Student Assistance

Language models such as ChatGPT exploit prompts to personalize the model's behavior and shape its outputs. A prompt [9] is defined as a set of instructions that programs the LLM to enhance and refine its capabilities. [14] review the most important patterns that can be used in the context of LLM. In our work, we focused on the categories of Interaction and Customization patterns. The Interaction category pattern allows the user to interact with the LLM in different ways, for example, by letting the LLM ask questions to the users to accomplish a task. The Customization category includes patterns that allow the user to limit the output to meeting specific types, and structures, such as indicating the LLM to play as a specific person. Although LLMs have shown drawbacks for students, such as cheating, plagiarism, and overreliance [15], the education field has also benefited from their use. ChatGPT [16], the most famous LLM nowadays, has made significant contributions to help students improve their competencies [17]. Students have been using ChatGPT to enhance their learning activities by quickly referencing it and utilizing it for self-study [18]. One of the recent works [19] uses ChatGPT to improve users' moral judgment and decisions. The experiments in the study ask ChatGPT questions such as "whether it is right to sacrifice one person's life to save those of five others" or other issues with the questions on abortion etc. and further prove that ChatGPT corrupts rather than improves the moral judgement of its users. These works motivated us to investigate whether ChatGPT can support students in classifying moral content on social media platforms.

3. Methods

3.1. Dataset

This study uses the publicly available Moral Foundations Twitter (MFT) corpus [20]. The dataset was annotated by at least three trained annotators, with the majority for each label considered $\geq 50\%$. The annotators considered morality as a set of vices and virtues according to Moral Foundation Theory [21], including care–harm, fairness–cheating, loyalty–betrayal, authority–subversion, and purity–degradation and provided the annotations for each tweet with the presence of the vice-virtues or non-moral. Specifically, each tweet was assigned a label indicating the presence or absence of each virtue and vice, or a label indicating that the tweet was non-moral. This resulted in a set of 11 labels for each tweet. However, to simplify the study and follow the existing work [22], we also consider the moral task as a binary classification task, where each tweet belongs either to the moral class (belonging to any moral vice/virtue category) or to the non-moral class. As a result, every tweet in the dataset falls into one of two classes: moral or non-moral. We consider the labels (moral/non-moral) of the dataset as ground truth and use these labels to evaluate the performance of students and ChatGPT in the binary task of moral classification. Due to time constraints concerning the execution time, we randomly selected 200 tweets from the publicly available MFT corpus and used these for experiments involving ChatGPT in our research. In addition, to compare students' performance with each other and with ChatGPT, we use the subset of 15 tweets for the moral classification task (moral/non-moral) as a part of the preliminary research for the pilot study to get a better understanding of the performance in qualitative terms.

3.2. Large Language Model- ChatGPT

We use ChatGPT (GPT-3.5-Turbo) model using Azure OpenAI[3] Service to classify the tweet text. Our current study only examines ChatGPT's performance in a zero-shot setting to explore its ability without training. However, we plan to analyze ChatGPT in one-shot and few-shot settings as part of our future research. Moreover, we utilized mixed prompt patterns by taking advantage of the ability of LLMs to mimic different personas and the capacity to adjust prompts by gathering more data for improved results. In detail, we leveraged the *persona pattern* and the *Flipped Interaction Pattern*. The Persona Pattern (PP) allows users to generate personalized outputs from a specific perspective. The purpose of this pattern is to give the LLM a specific character that guides its output and prioritizes details. We have also used the Flipped Interaction Pattern (FIP) to manage the conversation with ChatGPT to achieve the best prompt for our goal. In the FIP, after the user provides a brief description of the problem, the LLM takes charge of the conversation by asking specific questions that assist in creating an effective prompt capable of achieving the desired task.

3.3. Pilot Study Framework

A pilot study involving 67 adolescents from a secondary school was conducted for our study. Each participant was assigned a unique identifier and provided a link for direct access to the platform. The platform was designed as a basic user interface containing tweet text and moral/non-moral radio buttons to allow each participant to provide their labels for the binary moral classification task. Each student was given a similar set of 15 tweets in a random fashion order to classify the moral/non-moral content. We do not capture any personal data concerning the privacy of each individual. Further, all students were asked to provide informed consent to participate in agreement with the Data Protection Officer of our University.

4. Results

We first analyzed ChatGPT's results on 200 tweets dataset based on the different contexts. Then we compared the performance of students and ChatGPT on 15 tweets as a part of the pilot study. We use accuracy, macro variant of precision, recall, and F1 score as performance metrics.

4.1. Performance of ChatGPT based on different context (roles/personas)

From Table 1, it can be seen that the performance of ChatGPT is different when we specify different roles and personas in ChatGPT's prompt. It is observed that ChatGPT performs best

[3] https://learn.microsoft.com/en-us/azure/ai-services/openai/overview

Table 1
Performance Comparison of ChatGPT with different context (roles/personas) [200 tweets dataset]

Different ChatGPT's Context(ChatGPT acts as)	Precision	Recall	F1 score	Accuracy
Trained AI language model	75.37	73.62	73.87	74.88
Tutor for kids	74.60	74.74	74.65	74.87
Tutor for teenagers	73.55	73.60	73.57	73.84
Tutor for adults	**76.69**	**76.48**	**76.57**	**76.92**
Teenager	76.12	74.89	75.14	75.89

Table 2
Performance Comparison of ChatGPT and Pilot Study Students [15 tweets dataset]

Entity	Precision	Recall	F1 score	Accuracy
ChatGPT as Trained AI model	72.50	68.75	68.89	71.42
ChatGPT act as teenager	65.15	60.42	59.06	64.28
Total 67 Students (Avg/Std.dev)	61.92±13.27	60.69±11.96	59.72±12.51	61.04±11.82
Best-performing Student	86.60	86.60	86.60	86.66
Worst-performing Student	33.33	33.92	33.05	33.34

when it acts as a tutor for adults, resulting in 3.05% performance improvement in F1 score compared to other scenarios. We also noticed that ChatGPT does not perform well enough in its role as a tutor for children and adolescents with F1 scores of 74.65 and 73.57 respectively, even though these age groups are the ones that need more guidance and are the main focus groups for imparting education on the morality/non-morality of the content. As ChatGPT's ability to identify moral content varies across roles and performs poorly in the case of children and young people, this provided us with a future research direction to explore ChatGPT's generated explanations to explore any significant differences in helping kids and teenagers understand the moral/non-moral content.

4.2. Pilot Study Analysis: Performance Comparison of Students and ChatGPT

Table 2 shows that ChatGPT outperforms the average student with 71.00% and 5.31% improvement in accuracy in the context of the AI model and as a teenager. The student with 86.66 accuracy performs best among all students and ChatGPT, while 14 students out of 67 (student accuracy > 71.42) outperforms ChatGPT's best performance, indicating that some students are very knowledgeable about morality. The lowest performing student scores only 33.34 accuracy and 53.73% of students (36 out of 67) achieve less accuracy than the mean value, suggesting that sufficient guidance is urgently needed to help these teens grasp the overall concept of non-moral/moral content.

5. Limitations

In this study, we aim to investigate the extent to which ChatGPT can help students understand moral content in social media. However, there are certain limitations in our current work. Due to the disadvantage of the short length of tweets, students and ChatGPT may have trouble identifying whether a tweet is moral or non-moral content. Better training of students on moral content and a few-shot training scenario for ChatGPT could be beneficial to improve the performance of both students and ChatGPT in the classification task. While we use the GPT-3.5-Turbo model for the experiments, an investigation with a newer version of ChatGPT, such as GPT4 and other more advanced LLMs, could reveal better qualities of the LLMs' to classify the moral content. Time constraints limit the number of students and the dataset size for the pilot study. However, we plan to use a larger number of informative tweets from the ground-truth dataset to evaluate the performance of LLMs in identifying moral content.

6. Discussion and Conclusion

In this research, we conduct a preliminary study whose main objective is to investigate the capabilities of ChatGPT to guide students, especially children and adolescents, in identifying moral/non-moral content disseminated on social media platforms such as Twitter. We first analyze the performance of ChatGPT based on different roles and personas. We then conduct a pilot study in which students from the teenage group categorize tweets into moral/non-moral content. Our preliminary results suggest that ChatGPT performs better when it behaves as a tutor for adults rather than as a tutor for children and teenagers. Moreover, we find that ChatGPT outperforms the average student when acting as a trained AI model and teenager, but 20.89% of students still perform better than ChatGPT's best results. However, a large proportion of students (53.73%) perform even worse than the average scores for all students. Such results prompted us to analyze in more detail how students and ChatGPT understand the moral and non-moral labels separately. This first round of analysis has also led us to explore the explanations ChatGPT gives in the case of a tutor for children or young people to gain a deeper understanding of the impact of different personas on ChatGPT's evaluations. We will consider this as part of our future research. Hence, in this study, we perform investigations from a higher-level perspective, and we will look into in-depth details in our future works.

References

[1] S. Kaur, K. Kaur, R. Verma, et al., Impact of social media on mental health of adolescents, Journal of Pharmaceutical Negative Results (2022) 779–783.

[2] O. Araque, K. Kalimeri, L. Gatti, Sentiment and moral narratives during covid-19 (2020).

[3] R. Rezapour, L. Dinh, J. Diesner, Incorporating the measurement of moral foundations theory into analyzing stances on controversial topics, in: Proceedings of the 32nd ACM conference on hypertext and social media, 2021, pp. 177–188.

[4] M. Dehghani, K. Johnson, J. Hoover, E. Sagi, J. Garten, N. J. Parmar, S. Vaisey, R. Iliev, J. Graham, Purity homophily in social networks., Journal of Experimental Psychology: General 145 (2016) 366.

[5] D. Taibi, G. Fulantelli, V. Monteleone, D. Schicchi, L. Scifo, An innovative platform to promote social media literacy in school contexts, in: ECEL 2021 20th European Conference on e-Learning, Academic Conferences International limited, 2021, p. 460.

[6] C. Limongelli, D. Schicchi, D. Taibi, Enriching didactic similarity measures of concept maps by a deep learning based approach, in: 2021 25th International Conference Information Visualisation (IV), IEEE, 2021, pp. 261–266.

[7] D. Schicchi, B. Marino, D. Taibi, Exploring learning analytics on youtube: a tool to support students' interactions analysis, in: Proceedings of the 22nd International Conference on Computer Systems and Technologies, 2021, pp. 207–211.

[8] OpenAI, 2023. URL: https://chat.openai.com/.

[9] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, G. Neubig, Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, ACM Computing Surveys 55 (2023) 1–35.

[10] A. Upadhyaya, M. Fisichella, W. Nejdl, A multi-task model for emotion and offensive aided stance detection of climate change tweets, in: Proceedings of the ACM Web Conference 2023, 2023, pp. 3948–3958.

[11] A. Upadhyaya, M. Fisichella, W. Nejdl, A multi-task model for sentiment aided stance detection of climate change tweets, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 17, 2023, pp. 854–865.

[12] A. Upadhyaya, M. Fisichella, W. Nejdl, Towards sentiment and temporal aided stance detection of climate change tweets, Information Processing & Management 60 (2023).

[13] A. Upadhyaya, M. Fisichella, W. Nejdl, Intensity-valued emotions help stance detection of climate change twitter data, in: Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, ijcai.org, 2023, pp. 6246–6254.

[14] J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, D. C. Schmidt, A prompt pattern catalog to enhance prompt engineering with chatgpt, arXiv preprint arXiv:2302.11382 (2023).

[15] A. Tlili, B. Shehata, M. A. Adarkwah, A. Bozkurt, D. T. Hickey, R. Huang, B. Agyemang, What if the devil is my guardian angel: Chatgpt as a case study of using chatbots in education, Smart Learning Environments 10 (2023) 15.

[16] N. Fijačko, L. Gosak, G. Štiglic, C. T. Picard, M. J. Douma, Can chatgpt pass the life support exams without entering the american heart association course?, Resuscitation 185 (2023).

[17] C. K. Lo, What is the impact of chatgpt on education? a rapid review of the literature, Education Sciences 13 (2023) 410.

[18] S. Nisar, M. S. Aslam, Is chatgpt a good tool for t&cm students in studying pharmacology?, Available at SSRN 4324310 (2023).

[19] S. Krügel, A. Ostermaier, M. Uhl, Chatgpt's inconsistent moral advice influences users' judgment, Scientific Reports 13 (2023) 4569.

[20] J. Hoover, G. Portillo-Wightman, L. Yeh, S. Havaldar, A. M. Davani, Y. Lin, B. Kennedy, M. Atari, Z. Kamel, M. Mendlen, G. Moreno, C. Park, T. E. Chang, J. Chin, C. Leong, J. Y. Leung, A. Mirinjian, M. Dehghani, Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment, Social Psychological and Personality Science 11 (2020) 1057–1071.

[21] J. Graham, J. Haidt, S. Koleva, M. Motyl, R. Iyer, S. P. Wojcik, P. H. Ditto, Chapter two - moral foundations theory: The pragmatic validity of moral pluralism, volume 47 of *Advances in Experimental Social Psychology*, Academic Press, 2013, pp. 55–130.

[22] J. W. Burton, N. Cruz, U. Hahn, Reconsidering evidence of moral contagion in online social

networks, Nature Human Behaviour 5 (2021) 1629–1635.