SkOTaPA: A dataset for Skepticism Detection in Online Text after Persuasion Attempt

Smitha Muthya Sudheendra[1], Maral Abdollahi[2], Dongyeop Kang[1], Jisu Huh[2], Jaideep Srivastava[1]

[1]College of Science and Engineering, [2]Hubbard School of Journalism
University of Minnesota, Twin Cities
{muthy009, abdol022, dongyeop, jhuh, srivasta}@umn.edu

Abstract

Individuals often encounter persuasion attempts, during which a persuasion agent aims to persuade a target to change the target's emotions, beliefs, and behaviors. These persuasion attempts can be observed in various social settings, such as advertising, public health, political campaigns, and personal relationships. During these persuasion attempts, targets generally like to preserve their autonomy, so their responses often manifest in some form of resistance, like a skeptical reaction. In order to detect such skepticism in response to persuasion attempts on social media, we developed a corpus based on consumer psychology. In this paper, we consider one of the most prominent areas in which persuasion attempts unfold: social media influencer marketing. In this paper, we introduce the skepticism detection corpus, SkOTaPA, which was developed using multiple independent human annotations, and inter-coder reliability was evaluated with Krippendorff's alpha (0.709). We performed validity tests to show skepticism cannot be detected using other potential proxy variables like sentiment and sarcasm.

Keywords: Persuasion dialogue, skepticism detection corpus, social media, text analysis, influencer marketing, .

1. Introduction

Persuasion dialogue is where one participant in the dialogue tries to convince another participant to endorse some proposition or statement (Walton & Krabbe, 1995). Individuals often encounter these persuasion attempts, during which a persuasion agent (e.g., advertisers, influencers, doctors, etc.) aims to persuade a target by changing the target's emotions, beliefs, and behaviors. These persuasion attempts can be observed in various social settings such as advertising, sales, public health, politics, and personal relationships. One example of a persuasion attempt is when social media influencers try to sell a product to their followers by highlighting the features of the product. This is known as persuasive advertising, where the persuading agent appeals to the target using three main strategies: Ethos: ethics, credibility, and character, Logos: logic and reason, and Pathos: feelings and emotions.

Generally, when people encounter a persuasion attempt and recognize it as such, their perception changes. According to the Persuasion Knowledge Model (Friestad & Wright, 1994), recognizing a persuasion attempt changes the target's perceptions of the agent's intentions and, consequently, influences their responses. Since people like preserving their freedom and autonomy, they will likely resist persuasion threatening their autonomy. Consequently, a target's response to a persuasion attempt often manifests in one of three ways: skepticism, reactance, and inertia. Skepticism concentrates on ethos and logos, i.e., it focuses on credibility, logic, and evidence. Reactance is the negative reaction to the persuasion attempt, and inertia is when the target does not pay attention to the message being conveyed.

In this paper, we build a dataset to detect skepticism in text after a persuasion attempt. Hence, we consider one of the most prominent areas where persuasion attempts unfold – social media advertising and influencer marketing.

Previous skepticism detection research has mainly focused on vaccine skepticism (Beres et al., 2023; Kreutz & Daelemans, 2022). These computational methods focus on including other variables like network graphs and node information to identify skepticism. Our motivation for developing this corpus is to identify more general, non-topic-specific skepticism given a sentence without additional details on the speaker/writer based on consumer psychology.

In section 2, we discuss how we conceptualized skepticism. In section 3, based on the definition, we show the annotated dataset creation, followed by section 4, presenting the experimental results of using this dataset, and section 5, discussing the validity tests.

2. Conceptualizing Skepticism

Encyclopedia Britannica defines skepticism as "the attitude of doubting knowledge claims set forth in various areas." Similarly, the American Psychological Association Dictionary defines skepticism as "an attitude of questioning, disbelief, or doubt." In consumer psychology and advertising, skepticism is defined as "the tendency toward disbelief of advertising claims" (Obermiller & Spangenberg, 1998). Based on these definitions, we conceptualized skepticism in the context of persuasion attempts as doubt or a tendency toward disbelief about the truth of what is being said or shown. As this study focuses on the interaction between a persuasion agent and a target, it is set in the context of influencer marketing. During these persuasion episodes, the social media user does not believe what the influencer said in their post and expresses doubt about the truth of what is being said. Skepticism can be distinguished from other related concepts, such as cynicism, sarcasm, and irony. While cynicism refers to the tendency to disbelieve information in general, regardless of the

LREC-COLING 2024, pages 14871–14876

source (Kanter & Wortzel, 1985), skepticism during persuasion episodes refers to disbelief toward persuasion attempts made by persuasion agents.
Moreover, skepticism is different from sarcasm and irony, which express the opposite of what someone means. Both often use hyperboles to exaggerate what is being said, although the opposite is intended (Kunneman et al., 2015; Sykora, Elayan, & Jackson, 2020). Thus, sarcasm and irony are different from the concept of skepticism, which expresses disbelief or doubt toward claims made in persuasion episodes.
Skepticism is distinct as a concept, considering its antecedents and outcomes in the context of persuasion attempts. When a target encounters and recognizes a persuasion attempt, they realize that the persuasion agent employs different tactics to influence the target's responses through various psychological mediators, such as their attention, perceptions, emotions, or attitudes (Friestad & Wright, 1994). Skepticism alters how the persuasion episode unfolds and can lead to different consequences based on the target recognizing a persuasion attempt. For example, prior research on influencer marketing demonstrated that skeptical reactions in response to influencers' persuasion attempts often lead to lower purchase intentions and less positive attitudes toward the message, influencer, and brand (De Jans, Cauberghe, & Hudders, 2018; Van Reijmersdal & van Dam, 2020; Van Reijmersdal et al. 2016). Thus, we developed a dataset to detect skepticism to improve our understanding of how it is expressed in social media texts. Figure 1 shows an example of an interaction between a persuasion agent and a target.

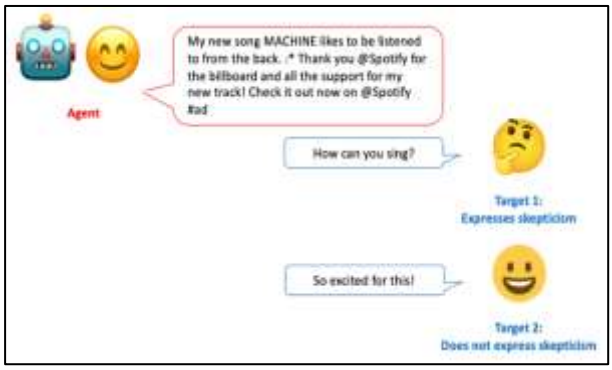

Figure 1: Example of an interaction between a persuasion agent and a target.

3. Annotating Skepticism

In this section, we apply our conceptualization of skepticism to the computational research domain for developing a training dataset to detect skepticism on social media.

3.1 Dataset

The dataset consisted of Instagram comments to social media influencers' posts since Instagram is one of the most popular platforms for influencer marketing. Moreover, users' comments on social media influencers' posts reflect targets' responses to persuasion attempts by agents and are therefore suited for this study. We used Instaloader (Graf & Koch-Kramer, 2019) to collect the captions and comments from Instagram. Social media is seeing a rise in Virtual influencers (VIs), especially in influencer marketing. Virtual influencers are replacing human influencers (HIs) due to human transgressions and technological evolutions. Hence, we considered comments to both HIs and VIs posts. We considered the top 25 HIs and 25 VIs who frequently engage in persuasion attempts using StarNgage and HypeAuditor. (Baklanov, 2022; StarNgage). We considered VIs since they tend to elicit doubt among social media users (Arsenyan and Mirowska, 2021). We chose comments to eleven VIs and three HIs who posted in English to create the dataset to be annotated. Since VIs did not have as many comments as HIs, we included more VIs than HIs to have a representative dataset. The English comments were randomly selected, and the dataset for annotation consisted of 9,818 comments.

3.2 Annotator motivation and experience

The annotators were selected based on the following criteria: 1. Regular social media users (using social media daily); 2. Familiar with and following social media influencers; 3. Have advanced consumer-level knowledge of social media marketing but less than industry expert-level knowledge. 4. Belong to the same age group that commonly uses Instagram, which, according to Pew Research Center (Auxier & Anderson, 2021), is 18-29. Additionally, we wanted to consider annotators without linguistic backgrounds as this corpus is built based on consumer psychology and not based on linguistic cues. The annotators worked regularly, were paid $12/hour, and reported weekly.

3.3 Instruction and annotation process

We developed our coding protocol for annotation based on the definition of skepticism from consumer psychology (Obermiller & Spangenberg, 1998), which is a seminal work in researching consumers' perceived skepticism in reaction to advertising messages. While research in the linguistics discipline provides insight into the language of skepticism, psychology research is more directly relevant to the purpose of our skepticism model development, which is focused on people's perceptions of skepticism in reaction to social media influencers' messages rather than identifying linguistic cues or structures for skeptical language. The coding protocol was tested and refined through pilot studies with human annotators, discussion of questions, misunderstandings, or reasons for differential coding outcomes from the annotators, and the annotation instructions and term definitions were refined to address the questions and discrepancies. For example, the initial protocol for the pilot study provided the coders with a straightforward definition of skepticism as follows: "The user expresses disbelief toward the post and/or influencer." compared to our final definition, skepticism refers to doubt or a tendency toward disbelief about the truth of what's being said or shown. In other words, the comment

shows that the user is not taking what the influencer said in their post at face value and expresses doubt about the truth".

Based on the initial rounds of pilot study and interactions with annotators, we identified various phrases and words associated with expressions of skepticism. These included using interrogative words, negations, if-clauses, and specific keywords conveying skepticism. For example, users frequently asked questions or phrased their skeptical responses as inquiries to question the authenticity of VIs. In some cases, they outwardly expressed their disbelief through negations, such as "I don't believe." We also considered keywords such as "fake" and "deceive" indicators of skepticism.

The recruited annotators underwent multiple training sessions to familiarize themselves with the task. During the training process, explicit instructions, along with instances of skeptical comments, such as "Are you a robot?" or "It's too good to be true." were provided to the annotators. Approximately 15 examples were provided to facilitate the annotators' comprehension of skeptical comments.

The coding process encompassed 14 rounds, with an average of approximately 400 comments coded in each round. These comments were a mix of textual content and emojis. The annotators were asked to label the comments as 'skeptical,' 'non-skeptical,' or 'unclear' if unsure. Throughout the annotation process, the annotators were also instructed to identify noteworthy keywords or phrases that aided in identifying skeptical comments.

Inter-coder reliability was consistently assessed using Krippendorff's alpha, yielding a value of 0.709 for the entire dataset. Comments annotated as 'unclear' by both annotators were dropped, and the final labels for these comments were chosen randomly when the annotators disagreed. The final annotated dataset[1] had 7,387 non-skeptical comments and 2,063 skeptical comments. These annotations were considered for the development of our prediction model.

4. Experiment

The final annotated dataset[1], consisting of 9,450 comments, was subjected to an 80-20 train-test split. Given the skewed nature of the dataset, we performed down-sampling on the training dataset. The final training dataset comprised 3,300 comments, while the test dataset comprised 1,890 comments. The test dataset comprised 413 skeptical comments and 1,477 non-skeptical comments. Table 1 shows the data description for the training and test dataset. Tables 2 and 3 show the data distribution for skeptical and not skeptical categories for training and test datasets, respectively.

We performed text classification using our training dataset on various baseline models to analyze our dataset. We considered the bert-base-uncased, bert-large-uncased (Devlin et al., 2019), RoBERTa-base, RoBERTa-large model (Liu et al., 2019), zero-shot gpt-3.5-turbo, 1-shot gpt-3.5-turbo, 3-shot gpt-3.5-turbo, and 10-shot gpt-3.5-turbo (OpenAI).

	training dataset	**test dataset**
Total comments	3300	1890
Unique tokens	5526	4297
Unique emojis	246	231
Comments with emojis	1074	714
Comments with questions	1610	814

Table 1: Data statistics for training and test dataset

	skeptical	**not skeptical**
Total comments	1650	1650
Unique tokens	3110	3599
Unique emojis	80	231
Comments with emojis	377	697
Comments with questions	934	676

Table 2: Data statistics per category for training dataset

	skeptical	**not skeptical**
Total comments	413	1477
Unique tokens	1239	3710
Unique emojis	46	216
Comments with emojis	95	619
Comments with questions	237	604

Table 3: Data statistics per category for test dataset

Each of the models considered for BERT and RoBERTa models were fine tuned for 5 epochs, learning rate of 1e-5 , batch size of 16, and optimized with AdamW optimizer (Loshchilov & Hutter, 2017), and sequence length of 512. The prompts provided for each GPT model was the definition of skepticism based on influencer marketing. We prompted the gpt models to explain and then provide the label since it has shown to provide better results (Wei et al, 2022).

4.1 Experimental Results

The models are evaluated based on the F1 score for each class and the macro-average. Table 4 shows the metrics for the models considered. The RoBERTa-large model performed the best, followed by BERT models and then 10-shot gpt-3.5-turbo. We investigated RoBERTa against 10-shot gpt-3.5-turbo further because they had a considerable difference compared to RoBERTa against the BERT models. Analyzing the predictions from 10-shot gpt-3.5 turbo and RoBERTa-large, we observed that the gpt-3.5 turbo model generally failed when the comments were short sentences (less than five words) or consisted mainly of emojis. Out of the misclassifications where

[1]Dataset : https://sites.google.com/umn.edu/skotapa-skepticism-dectection/home

Model	Not skeptical (F1 score)	Skeptical (F1 score)	Overall F1 score	macro-avg
bert-base-uncased	0.92	0.75	0.88	0.83
bert-large-uncased	0.92	0.77	0.88	0.84
RoBERTa-base	0.92	0.76	0.88	0.84
RoBERTa-large	**0.93**	**0.80**	**0.90**	**0.86**
zero- shot gpt-3.5-turbo	0.86	0.62	0.80	0.74
1- shot gpt-3.5-turbo	0.86	0.62	0.80	0.74
3- shot gpt-3.5-turbo	0.80	0.61	0.75	0.70
10- shot gpt-3.5-turbo	0.87	0.68	0.81	0.77

Table 4: Skepticism Detection Prediction Performance

10-shot gpt-3.5-turbo could not detect skepticism, 44.4% of the comments had less than five words and 33.33% had emojis in them. Out of the misclassifications where 10-shot gpt-3.5-turbo falsely detected skepticism, 38.8% of the comments had less than five words, and 17.10% contained emojis. There were only 13 instances where RoBERTa failed to detect skepticism when gpt-3.5-turbo detected skepticism correctly. Overall, RoBERTa-large performed better on the social media dataset. We included example predictions in Table 5. Analyzing the misclassifications further for RoBERTa-large, 6% of the comments were misclassified as skeptical contained interrogative words.

5. Validity Test

We performed validity tests to demonstrate that skepticism can be distinguished from other concepts like sarcasm, cynicism, and negative sentiment. We chose two of the existing and established concepts in the field of NLP, namely sentiment and sarcasm, to observe if they can be used as a proxy variable to detect skepticism. In order for us to use the concept (sentiment or sarcasm) as a proxy variable, we should be able to identify skepticism when we detect the proxy variable. If the statement expresses skepticism according to ground truth, the model should predict it as negative sentiment or sarcasm. We considered the SARC dataset (Khodak, Saunshi & Vodrahalli, 2017) to detect sarcasm and used the RoBERTa-large model for our analysis. The F1 score to detect skepticism was 0.15. We considered negative sentiment to be an indicator of skepticism. We used the IMDB dataset (Maas et al., 2011) for our analysis. The F1 score to detect skepticism was 0.48. Figure 2 depicts the heatmap for the proxy variables. The results show a negative correlation between sarcasm and skepticism, and sentiment and skepticism have a low correlation. Hence, they cannot be used as proxy variables for skepticism.

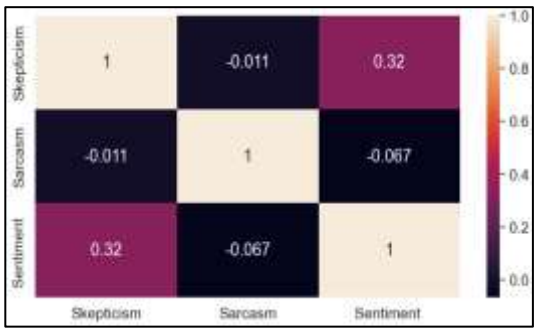

Figure 2: Heatmap of the correlation between the proxy variables

6. Limitations

In this paper, we have proposed a dataset that can be applied to detect skepticism after a persuasion attempt. However, this dataset is not without limitations. One of the main limitations is that the annotated dataset only consists of English comments. We plan to expand it to diverse annotators in future studies. Although this study uses real-world examples, we only considered one of the persuasion attempt scenarios. Hence, more domain-specific background work must be done to employ this study in other domains.

7. Conclusion

In this study, we investigated persuasion episodes between agents and targets and employed computational approaches to detect skepticism in

Comment	Human label	RoBERTa-large	10-shot gpt-3.5-turbo
I have a bad feeling about that😐	Skeptical	Not skeptical	Skeptical
Robotic beauty queen	Not skeptical	Skeptical	Not skeptical
This is scary	Not skeptical	Not skeptical	Skeptical
Your hands just exposed you😂	Skeptical	Skeptical	Not skeptical
Peep the wig 🤦🏾‍♀️👏🏾	Skeptical	Not skeptical	Not skeptical
I heard she doesn't respond	Not skeptical	Skeptical	Skeptical

Table 5: Examples of predictions by RoBERTa-large and gpt-3.5-turbo with ground truth label

social media text, taking the example of social marketing. First, we conceptualized skepticism from consumer psychology and advertising perspectives using the persuasion knowledge model and advertising skepticism research literature. Second, we introduced the dataset for skepticism. Then, we showed that the more recent and promising models like gpt-3.5-turbo cannot be used for identifying skepticism and we need to rely on human annotations and RoBERTa-large, but this can change soon given the rapid development of better large language models.

8. Ethical Consideration

In this paper, we have considered persuasion attempts between the agent and target within the context of influencer marketing. The need to discuss the ethical aspects is vital. The study relied on pre-existing interactions and employed a computational approach to understanding skepticism instead of survey-based methods. It is important to note that this method cannot be used in a dynamic real-time setting to manipulate the public as other variables influence each target. The primary objective of this study was to build a dataset to detect skepticism on social media after the interaction took place. The data used in the study is a public dataset, were as per terms of service, the identity of users in the comments was anonymized when the data was shared with the annotators.

9. References

American Psychological Association Dictionary (n.d.), "Skepticism," available at https://dictionary.apa.org/skepticism.

Auxier, B., and Monica Anderson (2021, April 7). Social media use in 2021. Pew Research Center: https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/

Baklanov, N. (2022, December 21). The Top Virtual Instagram Iinfluencers in 2022. HypeAuditor.com. https://hypeauditor.com/blog/the-top-virtual-instagram-influencers-in-2022/

Béres, F., Michaletzky, T. V., Csoma, R., & Benczúr, A. A. (2023). Network embedding aided vaccine skepticism detection. Applied Network Science, 8(1), 1-21.

Cambridge Dictionary (n.d.), "Sarcasm," available at https://dictionary.cambridge.org/us/dictionary/english/sarcasm.

Cortis, K., & Davis, B. (2021). Over a decade of social opinion mining: a systematic review. Artificial intelligence review, 54(7), 4873-4965.

De Jans, S., Cauberghe, V., & Hudders, L. (2018). How an advertising disclosure alerts young adolescents to sponsored vlogs: The moderating role of a peer-based advertising literacy intervention through an informational vlog. Journal of Advertising, 47(4), 309-325.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Find Influencers | StarNgage. (n.d.). StarNgage. https://starngage.com/app/global/influencer/search

Friestad, M., & Wright, P. (1994). The persuasion knowledge model: How people cope with persuasion attempts. Journal of consumer research, 21(1), 1-31.

Graf, A., & Koch-Kramer, A. (2019). Instaloader.

Kanter, D. L., & Wortzel, L. H. (1985). Cynicism and alienation as marketing considerations: some new ways to approach the female consumer. Journal of Consumer Marketing, 2(1), 5-15.

Khodak, M., Saunshi, N., & Vodrahalli, K. (2017). A large self-annotated corpus for sarcasm. arXiv preprint arXiv:1704.05579.

Kreutz, T., & Daelemans, W. (2022, June). Detecting Vaccine Skepticism on Twitter Using Heterogeneous Information Networks. In International Conference on Applications of Natural Language to Information Systems (pp. 370-381). Cham: Springer International Publishing.

Kreuz, R. J., & Glucksberg, S. (1989). How to be sarcastic: The echoic reminder theory of verbal irony. Journal of experimental psychology: General, 118(4), 374.

Kunneman, F., Liebrecht, C., Van Mulken, M., & Van den Bosch, A. (2015). Signaling sarcasm: From hyperbole to hashtag. Information Processing & Management, 51(4), 500-509.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.

Maas, A., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011, June). Learning word vectors for sentiment analysis. In Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies (pp. 142-150).

Obermiller, C., & Spangenberg, E. R. (1998). Development of a scale to measure consumer skepticism toward advertising. Journal of consumer psychology, 7(2), 159-186.

Priniski, J. H., & Holyoak, K. J. (2022). A darkening spring: How preexisting distrust shaped COVID-19 skepticism. PloS one, 17(1), e0263191.

Richard H. Popkin (n.d.), "Skepticism," Encyclopaedia Britannica, available at https://www.britannica.com/topic/skepticism/The-18th-century.

Sykora, M., Elayan, S., & Jackson, T. W. (2020). A qualitative analysis of sarcasm, irony and related# hashtags on Twitter. Big Data & Society, 7(2), 2053951720972735.

Van Reijmersdal, E. A., & van Dam, S. (2020). How age and disclosures of sponsored influencer videos affect adolescents' knowledge of persuasion and persuasion. Journal of youth and adolescence, 49(7), 1531-1544.

Van Reijmersdal, E. A., Fransen, M. L., Van Noort, G., Opree, S. J., Vandeberg, L., Reusch, S., ... & Boerman, S. C. (2016). Effects of disclosing sponsored content in blogs: How the use of resistance strategies mediates effects on

persuasion. American Behavioral Scientist, 60(12), 1458-1474.
Walton, D., & Krabbe, E. C. (1995). Commitment in dialogue: Basic concepts of interpersonal reasoning. SUNY press.
Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35, 24824-24837.