如何减少社交媒体上的攻击行为：基于北京大学树洞的案例研究

郭思含 2200017823　彭泳謙 2100017731　邬程灿 2300010746

1　引言

随着网络化的进程，社交媒体成为了人们日常生活中重要的一环，人们在社交媒体上花费大量的时间，而社交媒体上的内容也极大程度地影响着人们。在这种背景下，社交媒体上的攻击行为产生的负面影响，成为了一个值得关注的问题。本文将对北京大学树洞进行案例研究，以一般攻击模型（GAM）(Allen et al., 2018) 作为理论支撑，聚焦于：（1）发帖者本身语言的文明程度对攻击行为的影响、（2）评论者的攻击行为可能引起的锚定效应、（3）攻击约束对攻击行为的影响，三个方面进行研究，目的是从多个角度探究树洞上攻击行为的特征，以提出能够减少社交媒体中攻击行为的负面影响的意见。

2　相关工作

2.1　攻击欲望

2.1.1　一般攻击模型 (GAM)

一般攻击模型 (GAM) 是一个用于理解攻击行为的全面、综合的框架。它考虑了社会、认知、特质、发育和生物因素对攻击性的作用。GAM 的 “近似过程” 详细描述了人和情境因素如何影响认知、感受和唤醒，进而影响评估和决策过程，最终导向攻击性或非攻击性行为结果。

GAM 将攻击行为主要分为近端进程和远端进程。近端进程又分为三个阶段：输入、路线、结果。

输入阶段主要描述的是，人和情境因素如何通过影响第二阶段中当前的内部状态变量（即认知、情感和唤醒）来增加或减少攻击的可能性。具体地，情境因素是可能影响攻击是否发生的情境的各个方面。已有的研究发现许多情境因素会增加攻击的可能性。(Anderson and Carnagey, 2004; Bushman and Huesmann, 2010)

路线阶段关注的是，个人和情境因素通过哪些途径对评估和决策过程施加影响（从而影响攻击性或非攻击性结果）。个人和情境因素可以影响一个人的情感、认知和唤醒。这三个变量构成了个人的内部状态，它们的变化会影响攻击的可能性。

结果阶段关注评估和决策过程，以及攻击性或非攻击性的结果。在此阶段，个体评估当前情况并决定如何响应。选择行动后又会影响个体和情境因素，如此循环反复。

在研究中，我们参考一般攻击模型的作用机制，着重于研究输入阶段与路线阶段中的情境因素，利用控制变量的方法，解构情境因素在树洞攻击行为中所起到的作用。如图1。

2.2　攻击约束

2.2.1　规范性信念与共情

规范性信念是一种认知的自我调节机制，指导、评估恰当或可接受的行为 (Guerra et al., 1995; Huesmann and Guerra, 1997)，因此它是影响反应的重要因素。其中，共情在关于攻击的

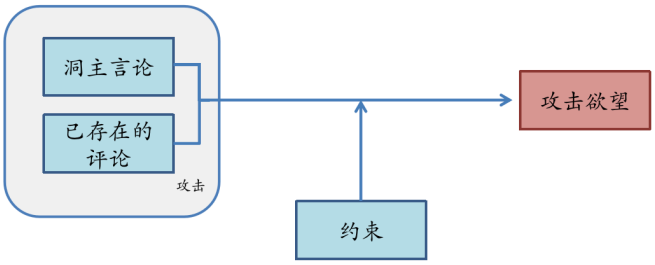

Figure 1: 洞主言论、已存在的评论、约束对攻击欲望的影响机制示意图。图中显示了洞主言论和已存在的评论通过攻击行为直接影响攻击欲望，而约束则作为一个调节变量，影响攻击欲望的高低。该示意图帮助解释了不同因素如何共同作用，最终影响个体的攻击欲望。

规范性信念中，起到了很重要的作用。共情是一种多方面的现象，包括认知共情和情感共情。认知共情是理解情绪状况并站在他人的角度思考的能力，而情感共情是间接体验他人的情绪的能力 (Davis, 1994; Hoffman, 1996)。有研究表明，高认知共情能力会导向高情感共情能力 (Batson et al., 2003)，并且有研究表明这两种能力与攻击性 (Miller and Eisenberg, 1988)、线下欺凌 (Mitsopoulou and Giovazolias, 2015) 和网络欺凌呈负相关。

在研究中，我们将影响攻击性的规范性信念构建成一种攻击约束，利用两段新闻报道材料，激起被试的共情，试图影响 GAM 中的路线阶段，探究在网络霸凌中，规范性信念对攻击性有何种程度的调节。

2.3 锚定效应

启发式是一个直观、快速、自动的系统 (Shiloh et al., 2002), 它“将评估概率和预测具体值的复杂任务简化为更简单的判断操作” (Tversky and Kahneman, 1974)，最早由 Herbert A. Simon 提出 (Simon, 1957)。而锚定效应，正是一个十分典型的启发式系统带来的效应，它作用于我们生活的方方面面，最早由 Paul Slovic 进行相关研究 (Slovic, 1967)。根据 Tversky and Kahneman (1974) 的研究，“锚定效应是指对决策者产生的不成比例的影响，使他们做出偏向于最初呈现的值的判断”。

在研究中，我们认为在社交媒体中的第一条评论至关重要，可能会起到锚定效应中锚定点的作用，从而极大程度地影响后续评论的整体风格。

3 方法

3.1 树洞实验

3.1.1 北京大学树洞

北京大学树洞是北京大学的一个匿名社交平台，是学生们分享自己的心情、困惑、秘密和故事的地方。这个平台让学生们能够在匿名的情况下表达自己，寻求支持和建议，缓解压力和焦虑。树洞的内容涵盖了学习、生活、情感、职业规划等各个方面。

- 洞主：发起新的帖子的用户

3.1.2 评论攻击性

0 分到 3 分，由实验者标注, 例：0：理解 dz，抱抱 dz；1：我觉得 dz 说的不好；2：dz 的想法真逆天；3：dz 真 sb，dz 脑子有问题。

问卷中也采用同一标准，设计具有不同攻击性的评论。以供被试选择希望发表的评论。

3.1.3 实验一

研究问题：在社交媒体中，发帖者使用的语言的文明程度和评论者使用的语言的攻击性的关系。

实验预期：发帖者语言文明程度越低，越会激起评论者的攻击欲望，最终使评论者的语言更具有攻击性。

自变量和因变量：实验的自变量是主试在树洞中发帖用词的文明程度，因变量是树洞中评论的攻击性。

实验步骤：主试先后在树洞发送一条内容相似、但使用词语的文明程度不同的树洞，词语依据是否文明分为：文明词语，不文明词语，讽刺词语 (讽刺词语指本身文明，但在语境中有不文明含义的词语)。在经过一定的时间间隔后，收集树洞内评论，并对评论的攻击性以及评论是支持洞主还是反对洞主进行分析。共发布 5 种内容，每条内容因用词文明程度不同，有 3 种版本，共 15 条树洞帖，收集 62 条评论。

3.1.4 实验二

研究问题：社交媒体上存在的攻击性言论，会引起锚定的效应，引发评论者更高的攻击水平。

实验预期：在存在攻击性评论的帖子中，被试的攻击水平有显著提高。

自变量和因变量：实验的自变量是已存在的评论的攻击性水平的高低，因变量是后续评论的攻击性，以及评论的攻击性是体现在支持洞主还是反对洞主。

实验步骤：主试先后在树洞中发送两条相同内容的帖子，之后助手分别在两条帖子中发送对洞主有攻击性的评论和对洞主没有攻击性的评论 (这两条评论均为该帖子的第一条言论)。在经过一定的时间间隔后，收集树洞内评论，并对评论的攻击性以及评论是支持洞主还是反对洞主进行分析。共发布 3 种内容，6 条树洞帖，收集 27 条评论。

3.2 问卷调查

3.2.1 实验三

研究问题：

- ◇ 在社交媒体中，发帖者言论的攻击性强弱，对评论者的攻击性的影响。
- ◇ 被试 (即评论者) 在阅读能激起同情心的新闻报道后，对发帖者攻击性的变化。
- ◇ 当发帖者言论的攻击性强弱不同时，被试在看完新闻报道后，对发帖者攻击性的变化是否不同。

衡量被试攻击水平的原理：统计被试对于具有不同攻击性语言的使用意愿，以及被试对于不同攻击性语言的攻击性的衡量。

实验预期：

- ◇ 发帖者言论的攻击性越强，评论者的攻击性越高。
- ◇ 在阅读新闻报道后，被试的攻击水平显著降低。
- ◇ 发帖者的攻击性越强，在阅读新闻报道后，被试的攻击水平降低得越不显著。

自变量和因变量：实验的自变量是发帖者的攻击性强弱，以及被试是否阅读了激起同情心的新闻报道。因变量是被试对发帖者的攻击性。

实验步骤：

- ◇ 先设定两种发帖内容，记为内容 A，内容 B，对于一种内容，有两种不同攻击性的版本，记为 A_{strong}，A_{weak}，B_{strong}，B_{weak}。
- ◇ 组间实验：
  - ∗ 将被试分为两组，两组中发帖者的攻击性强弱不同，于是记为强组和弱组，对于强组，一半的被试阅读 A_{strong}，一半的被试阅读 B_{strong}。
  - ∗ 对于弱组，一半的被试阅读 A_{weak}，一半的被试阅读 B_{weak}。
  - ∗ 在阅读发帖内容后，选择他们对于具有不同攻击性评论的发表意愿。
- ◇ 组内实验：
  - ∗ 两组被试均阅读一则新闻报道，再阅读另一个发贴内容。
  - ∗ 强组中，在组间实验中阅读 A_{strong} 的被试在看完材料后阅读 B_{strong}；在组间实验中阅读 B_{strong} 的被试在看完材料后阅读 A_{strong}。
  - ∗ 弱组中，被试在阅读发帖内容后，选择他们对于具有不同攻击性评论的发表意愿，并对比看完新闻报道后意愿的变化。
  - ∗ 一半被试先看 A 后看 B，另一半被试先看 B 后看 A，是为了避免材料本身不同对实验的影响。
- ◇ 共招募 40 名北大同学作为被试。

4 结果

4.1 实验一：洞主发帖文明程度不同对评论的影响

我们分析了当洞主发帖文明程度不同时，支持和反对洞主的评论占比，并计算支持洞主的评论的平均攻击性与反对洞主的评论的平均攻击性，根据发帖文明程度分为“文明组”“不文明组”“讽刺组”。

4.1.1 洞主发言文明程度对评论支持/反对的影响

支持和反对洞主言论占比，结果如图2。

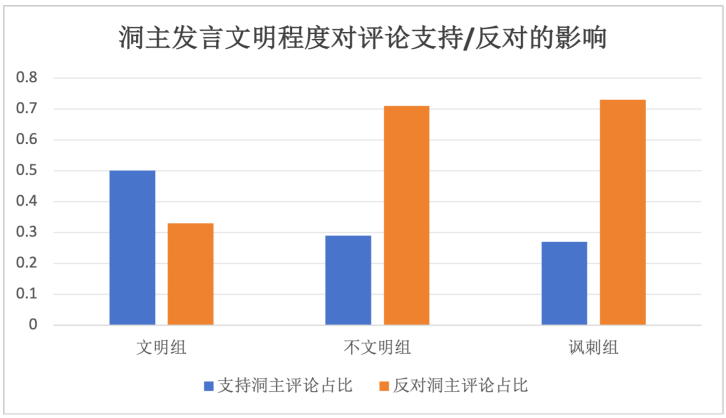

Figure 2: 图中展示了文明组、不文明组和讽刺组三组评论的支持和反对比例。蓝色柱状图表示支持洞主评论的比例，橙色柱状图表示反对洞主评论的比例。

结果显示：文明组中支持洞主的评论占 50%，反对洞主的评论占 33%，其余为中性态度的评论；不文明组中支持洞主的评论占 29%，反对洞主的评论占 71%；讽刺组中支持洞主的评论占 27%，反对洞主的评论占 73%。

这说明，当洞主发言文明时，评论者更倾向于发表支持洞主的言论；当洞主发言不文明或带有讽刺词汇时，评论者更倾向于发表反对洞主的言论。

4.1.2 洞主发言文明程度对评论攻击性的影响

支持洞主的评论的平均攻击性与反对洞主的评论的平均攻击性，结果如图3。

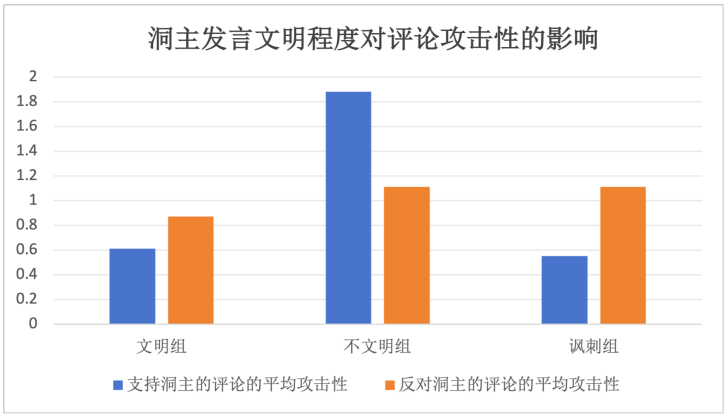

Figure 3: 图中展示了文明组、不文明组和讽刺组三组评论对支持和反对洞主的评论的攻击性。蓝色柱状图表示对支持洞主的评论的攻击性，橙色柱状图表示对反对洞主的评论的攻击性。

结果显示：文明组中支持洞主的评论平均攻击性为 0.61，反对洞主的评论平均攻击性为 0.87；不文明组中支持洞主的评论平均攻击性为 1.88，反对洞主的评论平均攻击性为 1.11；讽刺组中支持洞主的评论平均攻击性为 0.55，反对洞主的评论平均攻击性为 1.11。

这说明，当洞主发言文明时，无论是支持洞主的评论还是反对洞主的评论，攻击性都较低；当洞主发言不文明时，无论是支持洞主的评论还是反对洞主的评论，攻击性都较高。讽刺组反对洞主的攻击性略高于文明组，但并无明显差异。

当洞主发言不文明时，评论者的攻击性更强。并且支持洞主的评论，攻击性增强更明显；反对洞主的评论，攻击性增强不明显。

4.2 实验二：社交媒体上存在的攻击性言论引起的锚定效应

4.2.1 锚定效应与后续评论的态度

我们分析了当评论中已有支持/反对洞主的评论时，是否会引起锚定反应，导致后续的评论中“支持洞主的评论占比”和“反对洞主的评论占比”发生变化，结果如图4。

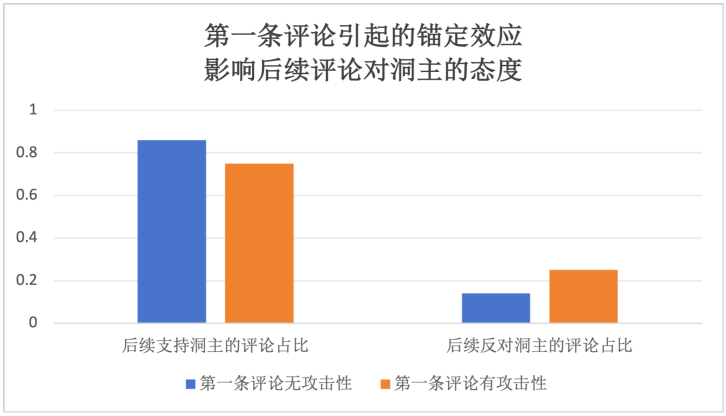

Figure 4: 图中展示了后续支持洞主的评论比例和后续反对洞主的评论比例，分别根据第一条评论是否具有攻击性进行分类。蓝色柱状图表示第一条评论无攻击性，橙色柱状图表示第一条评论有攻击性。

结果显示：当第一条评论为支持洞主时，后续支持洞主的评论占比为 86%，反对洞主的评论占比为 14%；当第一条评论为反对洞主时，后续支持洞主的评论占比为 75%，反对洞主的评论占比为 25%。

这说明，第一条评论引起的锚定效应，对后续支持/反对洞主的评论的占比有微弱影响。

4.2.2 锚定效应与后续评论的攻击性

我们分析了当评论中已有支持/反对洞主的评论时，是否会引起锚定反应，导致后续的评论攻击性发生变化，结果如图5。

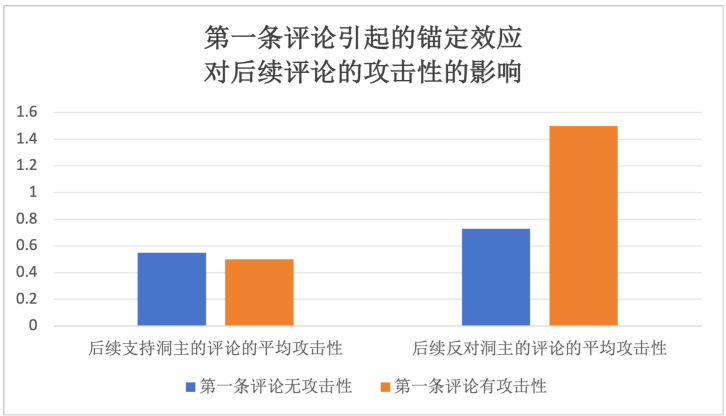

Figure 5: 图中展示了后续支持洞主的评论和后续反对洞主的评论的平均攻击性，分别根据第一条评论是否具有攻击性进行分类。蓝色柱状图表示第一条评论无攻击性，橙色柱状图表示第一条评论有攻击性。

结果显示：当第一条评论为支持洞主时，后续支持洞主的评论平均攻击性为 0.55，反对洞主的评论平均攻击性为 0.73；当第一条评论为反对洞主时，后续支持洞主的评论平均攻击性为 0.5，反对洞主的评论平均攻击性为 1.5。

这说明，第一条评论引起的锚定效应，对后续评论的攻击性有很大影响。

4.3 实验三：发帖者言论的攻击性强弱，与评论者的攻击性的关系

社交媒体中，发帖者言论的攻击性强弱，与评论者的攻击性的关系；被试在阅读能激起同情心的新闻报道后，与作为评论者使用攻击性语言的倾向的关系，如图6。

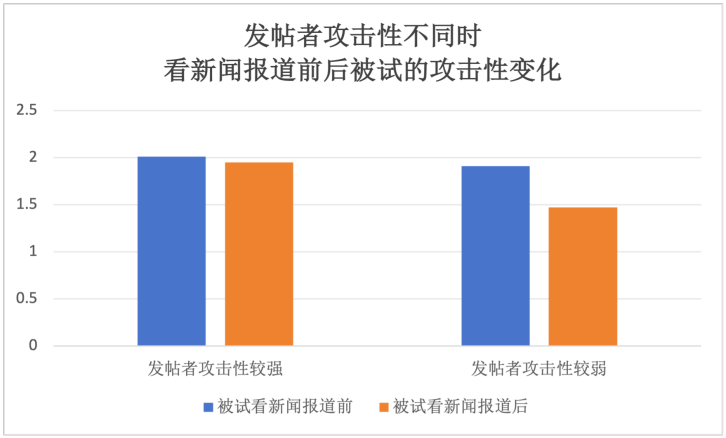

Figure 6: 图中展示了发帖者攻击性较强和发帖者攻击性较弱两种情况下，被试在阅读新闻报道前后的攻击性变化。蓝色柱状图表示被试在阅读新闻报道前的攻击性，橙色柱状图表示被试在阅读新闻报道后的攻击性。

结果显示：

- 阅读新闻报道前，强组的被试的平均攻击性为 2.01，弱组的平均攻击性为 1.91。
- 阅读新闻报道后，被试的攻击性显著降低，强组的被试的平均攻击性为 1.91，弱组的平均攻击性为 1.51。
- 强组在阅读新闻报道后，平均攻击性降低 0.10($p=0.409$)；弱组在阅读新闻报道后，平均攻击性降低 0.40($p<0.001$)。

这说明：

- 当发帖者攻击性较强时，被试的攻击性较强。
- 在阅读完新闻报道后，被试的攻击性均降低。
- 发帖者攻击性越强，在阅读完新闻报道后，攻击性下降得越不显著。

5 讨论

5.1 主要发现

洞主发言文明程度对于评论者的攻击性与是否支持洞主的倾向均有影响：当洞主发言文明时，评论者更倾向于发表支持洞主的言论；当洞主发言不文明或带有讽刺词汇时，评论者更倾向于发表反对洞主的言论。当洞主发言不文明时，评论者的攻击性更强。并且支持洞主的评论，攻击性增强得更明显；反对洞主的评论，攻击性增强得不明显。

第一条评论的锚定效应，对于后续评论的攻击性有很大影响，对于评论立场没有显著影响。

阅读能够激起同情心的新闻报道能够显著降低人们在网络发言中的攻击性，但发帖者的攻击性与此有交互作用：发帖者攻击性越强，在阅读完新闻报道后，攻击性下降得越不显著。

5.2 未来工作

首先，实验只探究了第一条评论对于被试的评论的锚定效应影响，但是现实中会有更多已有的评论，需要探究这些评论的立场、攻击性、数量等引起的从众效应。

其次，在树洞上发帖进行实验，被试仅限于本校大学生这一个群体，下一步的实验需要在更具代表性的平台进行，且尽量涉及更加多样的主题使得被试能够代表全体。

探究网络攻击性评论最终的目的是降低网络攻击性，营造健康的网络环境，减少悲剧的重现。因此，未来工作的一个重点是，如何将研究结果应用到实际生活中，使研究真正造福人们的生活。

6 致谢

感谢姜佟琳老师的悉心指导和帮助、谷承波学长在选题和实验设计方面的宝贵建议，使我们在研究过程中能够不断进步和完善。再次感谢老师和助教学长学姐在社会心理学课程中的悉心付出。

References

Allen, J. J., Anderson, C. A., and Bushman, B. J. (2018). The general aggression model. *Current opinion in psychology*, 19:75–80.

Anderson, C. A. and Carnagey, N. L. (2004). Violent evil and the general aggression model. *The social psychology of good and evil*, 168:192.

Batson, C. D., Lishner, D. A., Carpenter, A., Dulin, L., Harjusola-Webb, S., Stocks, E. L., Gale, S., Hassan, O., and Sampat, B. (2003). "... as you would have them do unto you" : Does imagining yourself in the other's place stimulate moral action? *Personality and social psychology bulletin*, 29(9):1190–1201.

Bushman, B. J. and Huesmann, L. R. (2010). Aggression. *Handbook of social psychology*.

Davis, M. (1994). Empathy: A social psychological approach. dubuque, ia: Wm. c. brown communications.

Guerra, N. G., Huesmann, L. R., and Hanish, L. (1995). The role of normative beliefs in children's social behavior.

Hoffman, M. L. (1996). Empathy and moral development. *The annual report of educational psychology in Japan*, 35:157–162.

Huesmann, L. R. and Guerra, N. G. (1997). Children's normative beliefs about aggression and aggressive behavior. *Journal of personality and social psychology*, 72(2):408.

Miller, P. A. and Eisenberg, N. (1988). The relation of empathy to aggressive and externalizing/antisocial behavior. *Psychological bulletin*, 103(3):324.

Mitsopoulou, E. and Giovazolias, T. (2015). Personality traits, empathy and bullying behavior: A meta-analytic approach. *Aggression and violent behavior*, 21:61–72.

Shiloh, S., Salton, E., and Sharabi, D. (2002). Individual differences in rational and intuitive thinking styles as predictors of heuristic responses and framing effects. *Personality and Individual Differences*, 32(3):415–429.

Simon, H. (1957). A behavioral model of rational choice. *Models of man, social and rational: Mathematical essays on rational human behavior in a social setting*, 6(1):241–260.

Slovic, P. (1967). The relative influence of probabilities and payoffs upon perceived risk of a gamble. *Psychonomic Science*, 9(4):223–224.

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157):1124–1131.