SafeLens Legal Guidelines

Introduction

SafeLens is dedicated to improving online content safety by fine-tuning foundational models across different modalities. Our system identifies and flags unsafe material, particularly content that may violate community guidelines or pose potential harm to users on social media platforms. We aim to develop agent-based models capable of monitoring and enforcing compliance with content safety regulations in real-time, helping detect harmful and unsafe content efficiently.

SafeLens Motto:
Guarding Digital Content
Ensuring Safe Online Content for All

Guiding Principles

SafeLens operates with a primary focus on ensuring the safety and well-being of online users by detecting and flagging unsafe content. Our system encourages users to contribute to our dataset by submitting both safe and unsafe content for fine-tuning our foundational models. However, it is essential to adhere to the following principles to maintain a respectful and responsible environment for content submission:

1. **Respect for Privacy**
   Users must refrain from uploading any content that violates the privacy of any individual. While we allow content categorized as "unsafe" (e.g., materials involving terrorism, violence, or explicit language), such content must not be directed at or target specific individuals in a harmful or defamatory manner. Content that involves personal vendettas, doxing, or exposure of private information without consent is strictly prohibited.

2. **Non-Personalized Unsafe Content**
   SafeLens acknowledges the need for unsafe content for research purposes to improve content safety measures. Therefore, we allow users to upload materials that are inherently unsafe, such as depictions of violence, terrorism, or harmful ideologies, provided they are presented in a general, non-targeted way. Any submission targeting specific persons or communities with the intent to harm, harass, or defame will not be accepted and will be flagged for removal.

3. **Legality and Ethical Responsibility**
   SafeLens differentiates between unsafe content for research purposes and illegal content. While we allow for the submission of unsafe content, we strictly prohibit the following:

   - **Illegal content**: This includes child exploitation, terrorism propaganda intended for criminal activity, and defamatory content against specific individuals.
   - **Hate speech and discrimination**: Any form of content inciting hate, violence, or discrimination against individuals or communities based on race, ethnicity, gender, religion, or other protected characteristics.
   - **Privacy Violations**: Personal information that could lead to the identification, harassment, or endangerment of an individual.

4. **Liability Waiver for Users**
   SafeLens is not responsible for any privacy violations or harm resulting from user-contributed data that breaches the privacy policies of individuals. While we take measures to monitor and flag content that violates our guidelines, users bear full responsibility for ensuring their submissions comply with SafeLens' rules regarding non-personalized unsafe content.

User Contributions

To enhance the effectiveness of SafeLens in identifying and flagging unsafe content, users can contribute to our dataset by uploading various types of media, including text, images, and audio. The system accepts both safe and unsafe content, provided it does not violate any legal or ethical guidelines outlined in this document.

Examples of Acceptable Unsafe Content for Research:

- **Terrorism-related media**: General materials that represent or describe acts of terrorism without targeting specific individuals or communities.
- **Hate speech detection**: Text or audio that may contain harmful ideologies, provided it does not promote real-world violence or is directed against individuals or groups.
- **Violence or graphic content**: Images or videos depicting violence in general but not intended to incite harm against specific persons or communities.

Examples of Unacceptable Content:

- **Targeted harassment**: Any form of media aimed at defaming, harming, or exposing an individual's personal life or private information.
- **Personal vendettas**: Content uploaded with the intent to harm someone with whom the user has a personal conflict.
- **Private information leakage**: Any media that exposes personal identifiers like addresses, contact details, or financial information of others.

Conclusion

SafeLens is committed to maintaining a safe, respectful, and responsible environment for users contributing to our research. By following these legal guidelines, you help us create a balanced dataset while protecting individual privacy and adhering to ethical standards. Should you encounter any issues or wish to report violations of these guidelines, please contact our support team for further assistance.

By contributing to SafeLens, users agree to abide by these legal guidelines and acknowledge their responsibility for the content they submit.

Ownership
SafeLens is created and operated by **Nitish Bhardwaj**. Any unauthorized use or reproduction of the content, system, or ideas without explicit permission from the founder will be subject to legal action.